

# **Virtual compound screening and SAR analysis: method development and practical applications in the design of new serine and cysteine protease inhibitors**

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Mihiret Tekeste Sisay

aus

Worka/Äthiopien

Bonn

June 2010

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen  
Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
2. Referent: Univ.-Prof. Dr. rer. nat. Michael Gütschow

Tag der Promotion: 01.10.2010

## Acknowledgments

I would like to express my deepest gratitude and appreciation to my supervisors Prof. Dr. Michael Gütschow and Prof. Dr. Jürgen Bajorath, for their continued guidance, encouragement and support both in my scientific and personal life throughout the course of this work.

I am especially indebted to all my colleges and colleges of my supervisors with whom I had successful scientific collaborations especially to Prof. Dr. G. König, Prof. Dr. T. Steinmetzer, Prof. Dr. S. Fustero, Dr. L. Peltason, Dr. M. Stirnberg, Dr. D. Stumpfe, L. Tan, T. J. Crisman, M. Frizler, E. Maurer and S. Hauptmann. I am also grateful to Dr. P. W. Elsinghorst, Dr. H.-G. Häcker and Dr. J. Batista for their helpful advices and discussions, and the rest of the members at the research groups of Prof. M. Gütschow and Prof. J. Bajorath for their friendship and all the good times I had.

I owe my deepest gratitude to my family for their endless love and am deeply grateful to Asnake Asegu for his all time encouragement and support. Especial thanks to Isa Schierstedt and her family for their encouragement and limitless support particularly to Dr. D. Schierstedt and M. Schierstedt.

Lastly, I offer my regards and blessings to all of those who in one way or another supported me.





*Dedicated to my parents*

ለአባቴ፡ተከሥተ፡ሲሳይ

ለእናቴ፡አበበች፡በቀለ

ለሚያደርግልኝ፡ሁሉ፡እግዚአብሔር፡የተመሰገነ፡ይሁን።



## Abstract

Virtual screening is an important tool in drug discovery that uses different computational methods to screen chemical databases for the identification of possible drug candidates. Most virtual screening methodologies are knowledge driven where the availability of information on either the nature of the target binding pocket or the type of ligand that is expected to bind is essential. In this regard, the information contained in X-ray crystal structures of protein-ligand complexes provides a detailed insight into the interactions between the protein and the ligand and opens the opportunity for further understanding of drug action and structure activity relationships at molecular level. Protein-ligand interaction information can be utilized to introduce target-specific interaction-based constraints in the design of focused combinatorial libraries. Furthermore, such information can also be directly transformed into structural interaction fingerprints and can be applied in virtual screening to analyze docking studies or filter compounds. However, the integration of protein-ligand interaction information into two-dimensional compound similarity searching is not fully explored. Therefore, novel methods are still required to efficiently utilize protein-ligand interaction information in two-dimensional ligand similarity searching. Furthermore, application of protein-ligand interaction information in the interpretation of SARs at the ligand level needs further exploration. Thus, utilization of three-dimensional protein ligand interaction information in virtual screening and SAR analysis was the major aim of this thesis. The thesis is presented in two major parts. In the first part, utilization of three-dimensional protein-ligand interaction information for the development of a new hybrid virtual screening method and analysis of the nature of SARs in analog series at molecular level is presented.

A new virtual screening hybrid methodology, termed the interaction annotated structural features, was introduced that assigns energy-based scores to two-dimensional substructures based on three-dimensional protein-ligand interaction information and utilize interaction-annotated features in virtual screening. Database molecules containing annotated fragments were assigned cumulative scores that serve as a measure of similarity to active reference compounds. In benchmark calculations on different high-throughput screening

data sets, the hybrid approach mostly performed better than conventional fragment-based two-dimensional fingerprint similarity searching and three-dimensional docking calculations.

On the other hand, to better understand how SAR discontinuity detected at the ligand level is reflected by three-dimensional protein-ligand interaction information, different compound series in combinatorial analog graphs were analyzed and substitution patterns that introduce activity cliffs were determined. The identified SAR determinants were then studied on the basis of three-dimensional ligand-target X-ray crystal complexes to enable a structural interpretation of SAR discontinuity and underlying activity cliffs. The analysis showed that many discontinuous SAR features extracted from combinatorial analog graphs can be directly associated with experimental three-dimensional receptor-ligand interactions. However, this was not always possible and some substitution site patterns that introduce significant SAR discontinuity in analog series cannot be explained in structural terms.

The second part of the thesis is focused on the application of different virtual screening methods for the identification of new cysteine and membrane-bound serine proteases inhibitors. In addition, molecular modeling studies were also applied to analyze the binding mode of cyclic peptide inhibitors.

Two major virtual screening campaigns were carried out to identify cathepsin K, cathepsin S and matriptase-2 inhibitors. While cathepsins K and S are cysteine proteases, matriptase-2 is a newly identified type II membrane-bound serine protease. These proteases are considered to be important current pharmaceutical targets due to their involvement in bone resorption, immune response and iron metabolism, respectively.

The first virtual screening application was focused at identification of dual cathepsin K and S inhibitors using a ligand-based compound mapping algorithm. By testing only 10 candidate compounds selected from a source database containing ~3.7 million molecules, two inhibitors of cathepsin K and S with new scaffolds were identified. Both inhibitors did not contain an electrophilic “warhead” that usually is present in most of the previously reported covalently interacting cathepsin inhibitors.

In a second study, through structure-based virtual screening in combination with similarity searching and knowledge-based compound design, two *N*-protected dipeptide amides containing a 4-amidinobenzylamide were identified as the first small molecule inhibitors of matriptase-2 with  $K_i$  values of 170 nM and 460 nM, respectively. An inhibitor of the closely related protease, matriptase-1 ( $K_i = 220$  nM) with more than 50-fold selectivity over matriptase-2 was also identified.

These newly identified inhibitors of the above proteases provide starting points for further chemical exploration of non-covalent cathepsins K and S inhibitors and non-peptidic matriptase-2 inhibitors.

Finally, three new cyanobacterial peptides, brunsvicamides A-C, were identified as selective inhibitors of human leukocyte elastase (HLE) through enzyme-based screening assays. Further molecular modeling studies were performed to analyze the possible binding mode of the cyclic peptides. The results showed that the cyclic peptides bind into the active site of HLE by forming several putative intermolecular interactions and mimicking an experimentally determined binding mode of a similar cyclic peptide.

# Contents

<b>1</b>	<b>General introduction</b>	<b>1</b>
<b>2</b>	<b>Integration of protein-ligand interaction information into 2D substructures for virtual screening</b>	<b>9</b>
2.1	Introduction	9
2.1.1	Protein-ligand interactions	10
2.1.2	Structural fingerprints	11
2.1.3	Similarity searching	13
2.2	Methodology	14
2.2.1	Data set	16
2.2.2	Atom-based scoring of protein-ligand interaction	18
2.2.3	Feature annotation	19
2.2.4	Calculations	19
2.3	Results and Discussion	20
2.3.1	The Interaction Annotated Structural Features (IASF) method	20
2.3.2	Analysis of IASF	20
2.3.3	Evaluation of the performance of IASF	22
2.3.4	Comparison of IASF and substructure searching	27
2.3.5	Analysis of annotated features	30
2.4	Summary	33
<b>3</b>	<b>Structural interpretation of activity cliffs revealed by systematic analysis of SARs in analog series</b>	<b>35</b>
3.1	Introduction	35
3.2	Methodology	37
3.2.1	Compound analog series and X-ray structures	37
3.2.2	Analysis of 3D protein-ligand interactions	38
3.2.3	R-group decomposition	38
3.2.4	Organization and analysis of analog series	39
3.2.5	SARI discontinuity scores	40
3.3	Results and Discussion	40
3.3.1	The CAG formulation	40
3.3.2	Patterns of SAR discontinuity and mapping of activity cliffs	42
3.4	Summary	52

---

<b>4</b>	<b>Identification of dual cathepsin K and S inhibitors</b>	<b>53</b>
4.1	Introduction	53
4.1.1	Cysteine proteases	54
4.1.2	Cathepsin K as a drug target	54
4.1.3	Cathepsin S as a drug target	55
4.1.4	Cathepsin inhibitors	55
4.2	Methodology	57
4.2.1	Data set and search strategy	57
4.3	Results and discussion	57
4.3.1	Binding mode analysis	60
4.4	Summary	61
<b>5</b>	<b>Identification of new matriptase-2 inhibitors</b>	<b>63</b>
5.1	Introduction	63
5.1.1	Type II membrane-bound serine proteases	63
5.1.2	The matriptase subfamily	64
5.1.3	Matriptase-1 as a drug target	64
5.1.4	Matriptase-2 as a drug target	65
5.2	Methodology	67
5.2.1	3D model of Matriptase-2	67
5.2.2	Virtual Screening calculations	67
5.2.3	Ligand-based virtual screening	68
5.2.4	Structure-based virtual screening	69
5.3	Results and discussion	70
5.3.1	Enzyme inhibition assays	73
5.3.2	Analysis of SAR and binding modes	74
5.4	Summary	78
<b>6</b>	<b>Inhibition and molecular modeling studies of brunsvicamides A - C against HLE</b>	<b>79</b>
6.1	Introduction	79
6.1.1	HLE as a drug target	79
6.1.2	Cyclic cyanobacterial peptides	80
6.1.3	The brunsvicamides	80
6.2	Methodology	81
6.3	Results and Discussion	82
6.3.1	Enzyme inhibition assays	82
6.3.2	Binding mode analysis	84
6.4	Summary	89

<b>7</b>	<b>Summary and conclusion</b>	<b>91</b>
<b>8</b>	<b>Appendices</b>	<b>93</b>
A	Software and Databases	93
B	Reference ligands from complex crystal structures	97
C	SAR tables	102
D	Screening data sets	109
E	Laboratory experimental details	113
<b>9</b>	<b>Bibliography</b>	<b>121</b>



## List of abbreviations

2D	Two dimensional
3D	Three dimensional
ADMET	absorption, distribution, metabolism, excretion and toxicity
CAG	Combinatorial analog graph
HIV	Human immunodeficiency virus reverse transcriptase
HLE	Human leukocyte elastase
HSP	heat shock protein 90
HTS	High-throughput screening
IASF	Interaction annotated structural features
JNK	c-jun N-terminal kinase 3
LB	Ligand-based
LBVS	Ligand-based virtual screening
MCS	Maximum common subgraph
MOE	Molecular operating environment
PDB	Protein data bank
PKA	Protein kinase A
PPE	Porcine pancreatic elastase
QSAR	Quantitative structure-activity relationship
SAR	Structure-activity relationship
SARI	Structure-activity relationship index
SB	Structure-based
SBVS	Structure-based virtual screening
Tc	Tanimoto coefficient
THR	Thrombin
TTSPs	Type II transmembrane serine proteases
VS	Virtual screening



# Chapter 1

## General introduction

Drug discovery is a highly complex and costly process, which requires integrated efforts involving innovation, information technologies, expertise, research and development investments, and management (Gershell, 2003). The economic pressure to bring drugs to the market has forced the pharmaceutical industry to embark on a complex drug discovery paradigm: searching for a new gene, a new target, new lead compounds and new drug candidates (Oprea, 2002). Discovery of innovative lead compounds is therefore, one of the key elements in a drug development project (Gershell, 2003; Kubinyi, 2007).

Over the past decade, the pharmaceutical industry has made large investments to establish high-throughput screening (HTS) technology for the identification of novel hits. HTS comprises the screening of large chemical libraries for activity against biological targets via the use of automation, miniaturized assays, and large-scale data analysis. HTS has indeed substantially contributed to the drug-discovery pipelines (Fox et al., 2006; Mayr and Bojanic, 2009). However, despite the great enthusiasm at the early stage, HTS has often also failed to identify active compounds that could be transformed into viable leads (Mestres, 2002; Stahl et al., 2002).

The drug discovery process has also been advanced by the use of virtual screening (VS) methods to identify new active compounds. These methods have emerged as alternative and complementary approaches to experimental HTS (Bajorath, 2002; Mestres, 2002; Langer et al., 2009). Moreover, *in silico* drug design supports decisions at different steps of the drug discovery process, such as the identification of a biomolecular target of therapeutic interest, selection or

the design of new active compounds, and their modification to improve potency and pharmacokinetic and pharmacodynamic properties. For VS diverse computational methods and tools are used to identify, rank and select candidate compounds in large compound libraries. The purpose is to reduce the magnitude and complexity of the screening problem, and to focus drug discovery and optimization efforts on the most promising molecules having desired properties and/or biological activity (Bajorath, 2002; Schneider and Böhm, 2002; Jorgensen, 2004).

In the initial phases of VS, filter criteria are often applied to eliminate compounds with undesired physicochemical properties such as the ‘rule of five’ (Lipinski et al., 2001) and/or presence of toxic or reactive structural fragments (Leeson et al., 2004; Chuprina et al., 2010). Since unfavorable *in vivo* properties (absorption, distribution, metabolism, excretion, and toxicity (ADMET)) of drug candidates frequently lead to attrition, increasing efforts are being made to define structural requirements for molecules to ultimately become drug candidates. (Lin et al., 2003; Kubinyi, 2007).

## **Structure- and ligand-based methods**

In VS, there are two fundamental approaches. These are structure-based VS (SBVS), which mostly involves molecular docking and requires knowledge of three-dimensional (3D) structure of the target binding site (Lyne, 2002; Kitchen et al., 2004; Shoichet, 2004; Ghosh et al., 2006), and ligand-based VS (LBVS), which includes fingerprint based methods, similarity searching, quantitative structure activity relationships (QSAR), comparative molecular field analysis (CoMFA), and pharmacophore methods but does not require information on the target structure.

LBVS methods extrapolate from known active compounds utilized as input information and aim at identifying structurally diverse compounds having similar biological activity. (Bender and Glen, 2002; Lengauer et al., 2004; Willett, 2006; Eckert and Bajorath, 2007; Taft et al., 2008; Langer et al., 2009; Geppert et al., 2010). It is in part based on the similarity property principle (Johnson and Maggiora, 1990) stating that structurally similar molecules are expected to have similar biological activity. Moreover, in chemical space representations, similar compounds usually map to similar regions, in other words, their intermolecular distance is expected be small, which is consistent with the similarity concept (Eckert and Bajorath, 2007).

Searching for compounds in databases that are similar to query molecules is one of the most widely applied LBVS approaches (Willett, 1998). Such similarity search methods are typically designed to capture structural

features and other properties of molecules in bit string format. Similarity search calculations are performed in “fingerprint space” (Eckert and Bajorath, 2007), which means that fingerprints are pre-calculated for query and database compounds and then quantitatively compared using various similarity metrics and coefficients such as the Tanimoto coefficient (Hert et al., 2004). Molecular fingerprints consist of various descriptors that are encoded as bit strings. Fingerprint overlap between test compounds is regarded as a measure of molecular similarity. Thus, if the chosen coefficient reaches a pre-defined threshold value, compared molecules are considered to be similar not only in structure but also in activity. In many fingerprint designs, a bit position accounts for a specific substructural feature or property and the bit is set on if this feature is present in the molecule. Furthermore, value ranges of other molecular descriptors (e.g., molecular weight or the number of hydrogen bond acceptors) can also be incrementally encoded as bit strings (Xue et al., 2003).

## Structure-based virtual screening

SBVS refers to the process of using the information contained in the 3D structure of a macromolecular target to design novel lead compounds that spatially fit the binding site forming energetically favorable intermolecular interactions (Lyne, 2002; Kitchen et al., 2004). Molecular docking (Kuntz et al., 1982) is the most widely used SBVS method that computationally searches for a ligand to fit the binding site of a protein target. It can be used as a primary hit identification tool when only structure of a target and its binding site is available or as a lead optimization tool when modifications to known active structures can quickly be tested in computer models before compound synthesis (Kitchen et al., 2004; Sperandio et al., 2006). To date, over 60 docking programs and more than 30 scoring functions have been introduced (Sperandio et al., 2006; Moitessier et al., 2008). The most widely used docking software tools include AutoDock (Goodsell and Olson 1990; Morris et al., 1998), DOCK (Kuntz et al. 1982), FlexX (Rarey et al. 1996) and Glide (Friesner et al. 2004). Each docking program relies on two complementary components: the proper positioning of the correct conformer of a ligand in the context of a binding site (posing) and its successful evaluation by a scoring function (scoring) (Kitchen et al., 2004; Sousa et al., 2006; Moitessier et al., 2008; Schulz-Gasch and Stahl, 2004; Leach et al., 2006).

Despite large efforts to improve the effectiveness of docking tools they often display inconsistent performance (Kitchen et al., 2004). In this context scoring represents the major problem, the reasons for this include (i) inadequate treatment of electrostatics, electronic polarization and aqueous desolvation, (ii) lack of accounting for entropy changes on accompanying binding, (iii)

insufficient ligand conformer sampling, and (v) the assumption of a rigid protein (Kitchen et al., 2004; Warren et al., 2006; Sousa et al., 2006; Moitessier et al., 2008; Kim and Skolnick; 2008). In recent years, it has been attempted to improve the performance of molecular docking and scoring by using new and advanced algorithms that take receptor plasticity, solvation and entropy into account (Schulz-Gasch and Stahl, 2004; Leach et al., 2006). The use of targeted scoring functions that are extended and recalibrated for a specific target or target class has been shown to further increase the performance of general scoring function (Seifert, 2009).

## **Integration of LBVS and SBVS methods**

LBVS and SBVS methods have been extensively utilized in drug design. Although the performance of VS methods is database and target protein dependent, several studies have shown that ligand based similarity searches are much more efficient than SBVS methods (Merlot et al., 2003; McGaughey et al., 2007; Hawkins et al., 2007; Tan et al., 2008). However, LBVS approaches often identify compounds that are structurally similar to the training set compounds, making it more difficult to identify novel chemotypes. On the other hand, as discussed above, docking methods are limited by the accuracy of the scoring function (Kitchen et al., 2004; Sperandio et al., 2006; Kim and Skolnick; 2008). Therefore, approaches to efficiently combine the two methods are highly desired because both ligand similarity and binding site information can be simultaneously utilized to maximize VS information content.

For example, LBVS and SBVS can be used sequentially where ligand-based screening methods are initially used to reduce the size of a large compound database to a reasonable number for structure-based design by molecular docking (Kitchen et al., 2004). It is also possible to integrate ligand information with docking where candidate compounds from similarity searching are instantly subjected to docking, enabling pre-computed ligand similarities to be incorporated into the docking and scoring process (Vidal et al., 2006). Another hybrid approach, that uses ligand-based scoring, compares the shape and chemical features of a candidate compound with a 3D reference ligand (Zavodszky et al., 2009).

In this thesis, both LBVS and SBVS methods were applied to identify new inhibitors of selected potential drug targets, including two cysteine proteases and a membrane bound serine protease. Two major VS campaigns were carried out to identify cathepsin K and S dual inhibitors and matriptase-2 inhibitors. These proteases are considered to be important current pharmaceutical targets due to their involvement in bone resorption (Blair and

Athanasou, 2004; Stoch and Wagner, 2008), immune response (Driessen et al., 1999; Honey and Rudensky, 2003) and iron metabolism (Du et al., 2008; Finberg et al., 2008; Melis et al., 2008; Ramsay et al., 2009), respectively.

## **Understanding and utilizing of protein ligand interactions**

Understanding the guiding principles of interactions between a target protein and a ligand is of paramount importance in drug design (Böhm and Klebe, 1996; Böhm and Schneider, 2003). Both strength and specificity of protein-ligand interaction arise from the accumulation of many forces between the ligand and the protein target such as electrostatic interaction, hydrogen bonding, van der Waals interactions, cation- $\pi$  interactions, metal complexation, and hydrophobic effects (Böhm and Klebe, 1996; Böhm and Schneider, 2003; Whitesides and Krishnamurthy, 2005). Knowledge of the 3D structure of a protein target and its ligand-binding site is a fundamental step in understanding the properties and function of the protein and molecular recognition mechanisms (Böhm and Schneider, 2003). In this regard, the crystal structure of a ligand bound to a protein target provides detailed insights into intermolecular interactions. The 3D interaction information can also be extracted and utilized to improve the performance of conventional two-dimensional (2D) fingerprint-based similarity search methods. Attempts have recently been made to directly and indirectly capture protein-ligand interaction information extracted from 3D complex structures for application in VS (Tan et al., 2008b; Tan and Bajorath, 2009; Deng, et al., 2004, Kelly and Mancera, 2004, Chuaqui et al., 2005; Singh et al., 2006; Marcou and Rognan, 2007).

In this thesis, it is also described how ligand-target interaction information can be utilized to complement conventional 2D similarity search methods. A new methodology is introduced to extract 3D interaction information on a per-atom basis and use it for scoring. Annotated substructures can then be applied in VS (Crisman et al., 2008).

## **Structure-activity relationships**

The relationship between chemical structure and biological activity of molecules, termed structure-activity relationship (SAR), is one of the most important aspects in drug design. The observed specific biological activity of a ligand, to the most extent, is governed by 3D intermolecular interactions with a macromolecular target. Therefore, small-molecule SAR studies should also take knowledge of specific protein-ligand interactions into consideration (Bender and Glen, 2002). As discussed above, protein-ligand binding is determined by the existence of specific intermolecular interactions of different chemical nature,

shape complementarities and other entropic effects (Böhm and Klebe, 1996; Bender and Glen, 2002; Eckert and Bajorath, 2007). It is also known that a single interaction, such as a hydrogen bond, can dramatically alter the selectivity and/or potency of a compound. This can account for abrupt changes in the biological response to minor chemical modifications of active compounds in analogs. Hence, information obtained from X-ray crystal structures of ligand-target complexes of analogs might provide critical information for structural interpretation of the nature of SARs.

In this thesis, to better understand SAR discontinuity at the molecular level of detail, different compound series were studied in a data structure termed combinatorial analog graphs and substitution patterns that introduce activity cliffs of varying magnitude were determined. So identified SAR determinants were then analyzed on the basis of 3D ligand-target X-ray crystal structures to enable a structural interpretation of SAR discontinuity and underlying activity cliffs.



## 1.2. Goals and approaches

The major focal point of this thesis project have been the development of new VS methods, analysis of SARs in analog series, and practical application of virtual screening methodologies for the identification of new inhibitors of selected cysteine proteases and a membrane-bound serine protease. The major goals of the thesis include:

*Goal 1:* Utilization of protein-ligand interaction information to develop new ligand-target interaction-based 2D similarity searching method;

*Goal 2:* Identification of SAR determinants in analog series using combinatorial analog graphs and subsequent analysis of activity cliff determinants based on 3D ligand-target interaction information;

*Goal 3:* Practical application of virtual screening methods for the identification of new inhibitors of selected pharmaceutical targets namely, cathepsin K, cathepsin S and matriptase-2; and

*Goal 4:* Enzyme inhibition assay and molecular modeling studies on three cyclic peptides against human leukocyte elastase and related enzymes.

## 1.3. Thesis Outline

The thesis is presented in two parts. The first part reports the development of a new VS method and analysis of SARs in analog series using 3D protein-ligand interaction information. This part is divided into two chapters, *chapters 2 and 3*.

*Chapter 2* focuses on the methodology and results of a new hybrid 2D/3D virtual screening approach termed here the interaction annotated structural features (IASF) method. Details on method development and performance evaluation in real HTS datasets are presented.

In *Chapter 3*, the nature of SAR discontinuity in analog series is analyzed at the level of protein-ligand interaction.

The second part of the thesis focuses on the practical application of different VS methods for the identification of new inhibitors of selected novel protein targets. The results are summarized in *chapters 4 and 5*

*Chapter 4* presents application of a LBVS study to identify dual inhibitors of two cysteine proteases, cathepsins K and S.

In *Chapter 5*, details of the results of combined LBVS and SBVS efforts to identify new inhibitors of human matriptase-2 are reported.

*Chapter 6* reports screening results and molecular modeling studies on three analogous cyanobacterial cyclic peptides against human leukocyte elastase.

Finally, *Chapter 7* summarizes the major results and presents general conclusions of the dissertation.

# Chapter 2

## Integration of protein-ligand interaction information into 2D substructures for virtual screening

The first part of the thesis, presented in chapters 2 and 3, reports studies carried out on VS method development and analysis of determinants of structure activity relationship discontinuities identified by combinatorial analogue graphs utilizing 3D protein-ligand interaction information. This chapter summarizes the development of a new VS methodology, the interaction annotated structural features (IASF) method (Crisman et al., 2008) and the next chapter presents SB interpretation of determinants of SAR discontinuities (Sisay et al., 2009a).

### 2.1 Introduction

A veritable plethora of chemical descriptors have been devised over the years, and, intuitively, one would think that those reflecting the 3D properties of a molecule would be more effective because the binding of a small molecule to a protein target is a 3D-dependent event. However, this is not always the case because most 3D-based methods, including molecular docking, rely on computational simplifications that significantly reduce the efficacy of the analytical process thereby making them less effective than 2D-based techniques, such as those that employ chemical substructures (Merlot et al., 2003). Similar trend has been observed when comparing molecular docking with simple 2D-based similarity methods (Tan et al., 2008) or with 3D ligand centric shape

matching (Hawkins et al., 2007, McGaughey et al., 2007). Nevertheless, since binding of a ligand to a target protein (protein-ligand interaction) is certainly a 3D event, which involves the 3D chemical and geometric complementarities of surfaces of both interacting molecules, efficient application of a 3D method might be more appropriate for defining molecular similarity.

Taking this fact in to consideration, devising novel methods that indirectly combine or encode 3D protein-ligand interaction information into 2D similarity search approaches would maximize on the performance of 2D-based screening methods by focusing on matching of interacting substructural features. It can potentially be used to retrieve compounds with low structural similarity but contain the most important interacting substructural features that are critical for the successful interaction of the ligand with the target protein.

### **2.1.1 Protein-ligand interactions**

The crystal structure of a protein-ligand complex provides a detailed insight into the interactions between the protein and the ligand. The interaction information can be used to identify where the ligand can be changed to improve the activity, physicochemical or ADMET properties of the compound, by identifying which parts of the compound are important for activity and which parts can be altered without affecting ligand binding. This can be further extended to map the specific substructures of the ligand which account most of the energetic contributions for binding. Such substructural features derived from protein-ligand interaction can be used to prioritize substructures in 2D similarity search methods. Therefore, combining or encoding 3D protein-ligand interaction information into 2D-based search methods will allow the use of valuable information in a more quantitative and objective way. Such combinations of computational approaches have been utilized in different cases to augment the capabilities of the individual methods such as SB pharmacophore searches derived from protein-ligand X-ray crystal structures (Griffith et al., 2005) and application of docking derived interaction information to ligand similarity searching (Briem and Kuntz, 1996).

Based on this basic idea, an effort was made to develop an alternative approach to conventional 2D fragment mapping that takes 3D protein-ligand interaction information into account. Interaction fingerprints have previously been reported that encode specific protein-ligand interaction information (Deng et al., 2004; Kelly and Mancera, 2004; Chuaqui et al., 2005; Singh et al., 2006; Marcou and Rognan, 2007; Pérez-Nueno et al., 2009). However, different from such representations, the aim of this work was designing a ligand-centric fragment approach that utilizes ensembles of 2D structural features and quantitatively annotates them with protein-ligand interaction information. For a

set of active reference molecules, interactions in crystallographic protein-ligand complexes are scored on a per-atom basis using an energy function and the atom-based scores are added to substructural features calculated for ligands using the extended connectivity fingerprints (ECFPs)<sup>1</sup>. The so derived sets of annotated substructures are used for database searching and subsequent ranking. In this chapter, a new VS methodology, the Interaction Annotated Structural Features (IASF) method, which utilizes protein-ligand interaction information to evaluate or rank database compounds based on 2D structural features, is reported.

### 2.1.2 Structural fingerprints

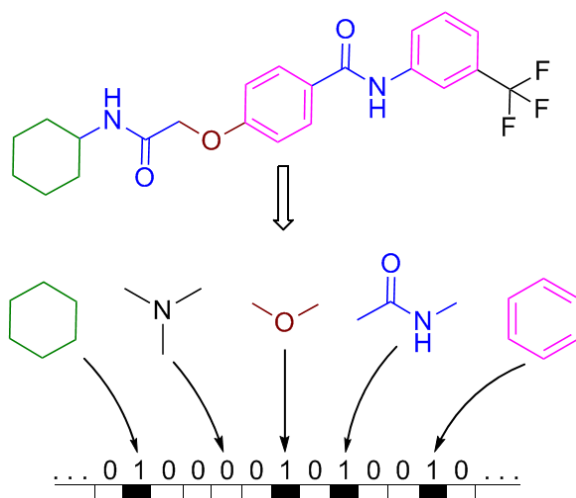
The use of molecular substructures or fragments has been extensively applied in computer-aided drug design and chemoinformatics, and has a long history in chemical and pharmaceutical research (Xue and Bajorath, 2000; Merlot et al., 2003; Mauser and Stahl, 2007; Hajduk and Greer, 2007; Congreve et al., 2008). Fragments are among the most popular molecular descriptors for compound clustering, in studying structure-activity relationships or database searching (Xue and Bajorath, 2000; Merlot et al., 2003) and are frequently used for 2D similarity-based or 3D SB *de novo* compound design (Mauser and Stahl, 2007). Furthermore, they often serve as building blocks for SB ligand design (Hajduk and Greer, 2007) and other fragment linking schemes (Crisman et al., 2008). Substructural fragments can also be used in the identification of new building blocks rich in biological motifs that can be utilized for synthesis planning (Lewell et al., 1998).

Fingerprints are bit-string representations of molecular substructures and properties. They represent a particularly popular format of fragment descriptors and are widely applied in compound clustering (Brown and Martin, 1998) and similarity searching (Willett et al., 1998; Willett, 2006). Structural fingerprints designed for similarity searching can essentially be either hashed connectivity pathways, structural dictionary-based or layered atom environments (Eckert and Bajorath, 2007). In the context of this work, two categories were considered; dictionary-based (keyed fingerprints) and layered atom environments (feature collections). Keyed fingerprints typically have a fixed format where each bit position is associated with a particular fragment whose presence or absence in a test molecule is monitored. They register the presence or absence of a substructure by a 1 or 0 in the corresponding position in a bit string (Figure 2.1). Pioneering developments of such fragment dictionary-based fingerprints include MACCS structural keys (Durant et al., 2002) or BCI fingerprints

---

<sup>1</sup> Fingerprint methods, software and databases used in this thesis are provided in Appendix A

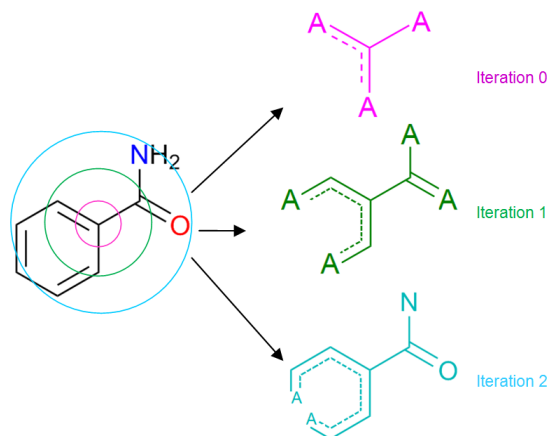
(Barnard and Downs, 1997). Recently, keyed fingerprints having a more variable format have also been introduced that encode varying numbers of compound class characteristic substructures (Batista and Bajorath, 2007).



**Figure 2.1: Keyed fingerprints.** The substructural features of a molecule are coded as bits in a fixed length bit-string and each bit position monitors the presence or absence of a specific substructural feature (such as a carbonyl, an amide bond, tertiary nitrogen or a saturated or unsaturated ring system) by recording as 1 or 0, respectively.

The second class of structural fingerprints, termed here feature collections, essentially represent layered atom environments that are systematically calculated for test molecules and recorded as individual strings or features. Since the number of accessible atom environments (and thus strings) can become exceedingly large, environments cannot be assigned to pre-defined bit positions, but are stored as individual sets of features. Thus, in contrast to keyed fingerprints, varying numbers of strings are generated for different test molecules. Similarity measures are then defined using set operators. For example, the intersection between two sets would correspond to the number of shared “1” bit positions in a keyed fingerprint. Pioneering designs of structural atom environment fingerprints include Molprint2D (Bender et al., 2004a; Bender et al., 2004b) or Scitegic’s Extended Connectivity Fingerprints (ECFPs) implemented in the Pipeline Pilot software.

For the generation of ECFPs, a code is assigned to each non-hydrogen atom consisting of its mass, valence, atom charge, atom type and the number of bonds to other atoms (to hydrogens and non-hydrogens). The atom code is combined with bond information and codes of neighboring atoms through a hashing procedure; features are sampled iteratively until a pre-defined bond diameter (layer) is reached (Figure 2.2). These features represent substructures and are recorded as large integers for each molecule and duplicates are removed. Information gain diminishes at higher iteration.



**Figure 2.2: Generation of ECFP fingerprints.** Initially, a code is assigned to each non-hydrogen atom consisting of its mass, atom charge, bond-type and atom-type. The atom code is combined with bond information and codes of neighboring atoms through a hashing procedure. Following that, features are sampled iteratively until a pre-defined bond diameter (layer). Arbitrary iteration layers and the resulting features are shown.

### 2.1.3 Similarity searching

Similarity searching using fragment-type fingerprints is a typical LB 2D mapping procedure. Fingerprints of reference molecule(s) are calculated and compared to corresponding fingerprints of database compounds; fingerprint overlap (corresponding to the number of shared fragments) is quantified via a similarity coefficient. Various similarity metric exist that return a score indicating the level of similarity between two molecules under comparison (Willett, 1998; Eckert and Bajorath, 2007) termed similarity coefficients. The Tanimoto coefficient (Tc) (Willett, 1998; Hert et al., 2004a; Hert et al., 2004b) is often used as a similarity measure and is calculated by taking the ratio between intersection and union of the bit sets of features between two compounds. Considering two molecules A and B, if  $a$  is the number of features present in A (bits set on in A), if  $b$  is the number of features present in B (bits set on in B), and if  $c$  is the number of features common to both molecules (bits set on in both A and B), the Tc for the two molecules is given as:

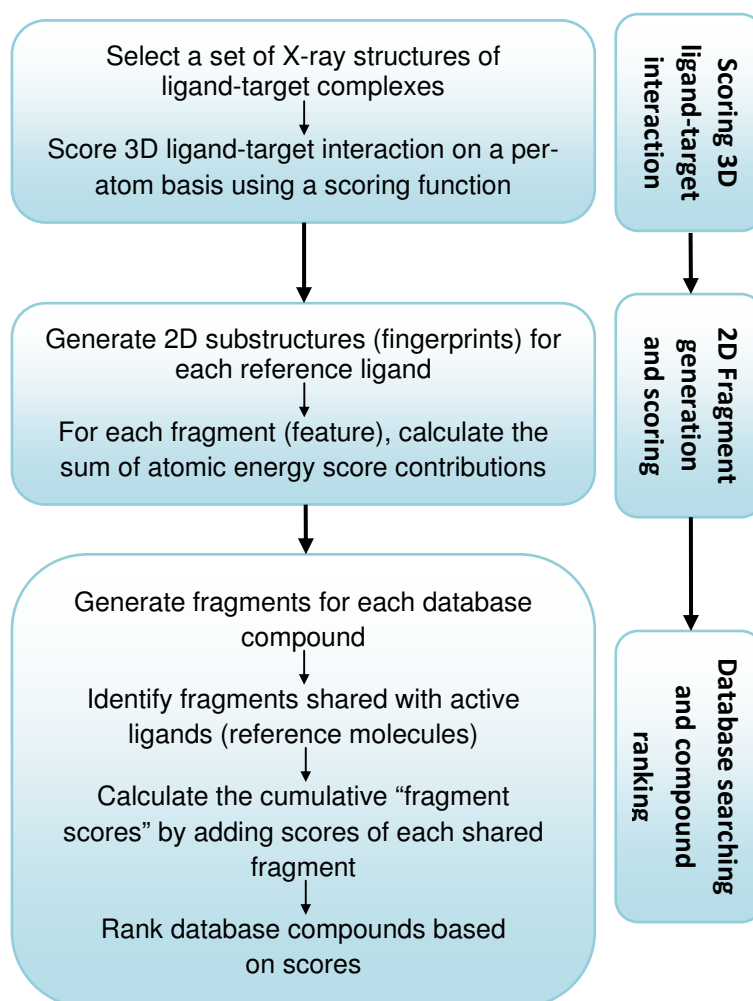
$$Tc(A, B) = \frac{c}{a+b-c} \quad (2.1)$$

It has a range between 0 (indicating dissimilar molecules) and 1 (similar molecules) but does not necessarily mean the molecules are identical.

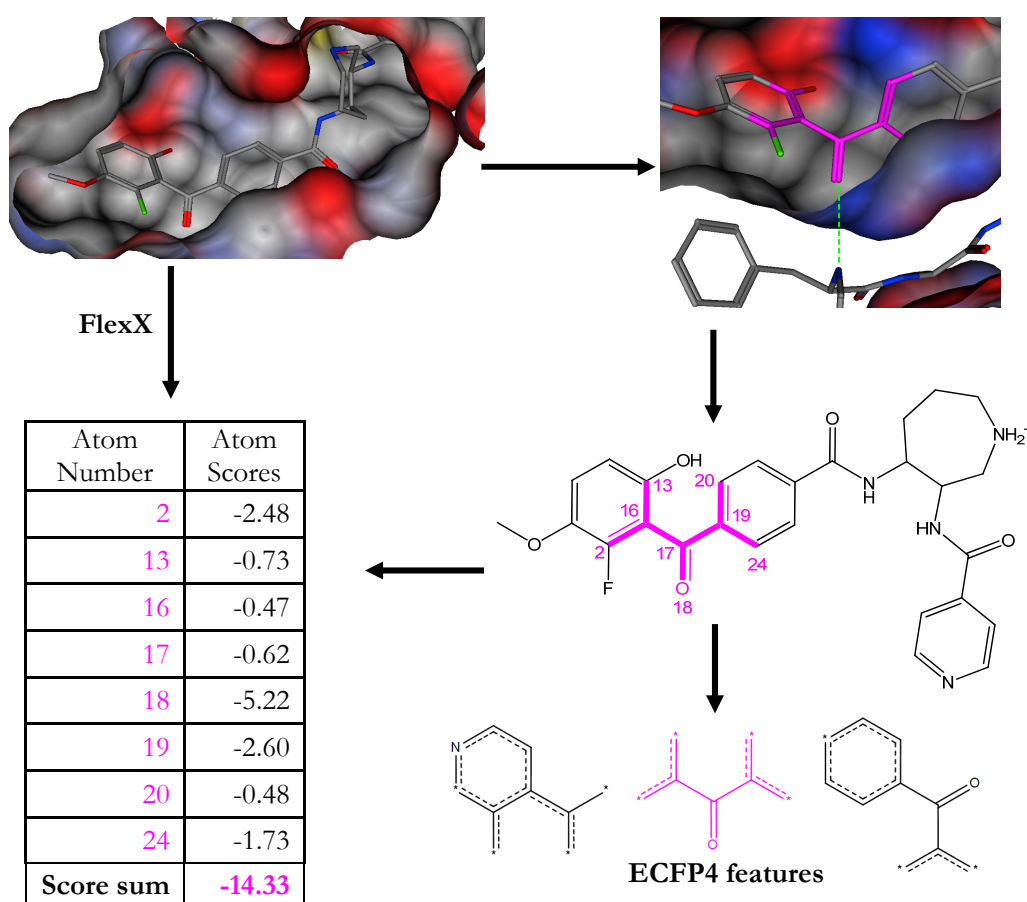
## 2.2 Methodology

The different steps involved in IASF calculations are summarized in Figure 2.3 and its key aspects are highlighted in Figure 2.4. A set of ligands available in crystal structures of protein-ligand complexes were selected and each protein-ligand complex was scored using the FlexX scoring function (Rarey et al., 1996; Rarey et al., 1997). The scoring gives detailed energy values of individual heavy atoms in a ligand. In parallel, ECFP4 structural features were generated for the selected ligands and were annotated with energy scores reflecting atomic contributions to interaction energies obtained from the FlexX per-atom scores. Database compounds were screened for corresponding features and obtain cumulative energy scores based on the features they contain. This produces a compound ranking by increasing cumulative scores that are utilized as a measure of similarity between active and database compounds (pipeline pilot script generation and data analysis was performed together with T. J. Crisman). It would be considerably more difficult to determine cumulative score cutoff values as an indicator of activity, due to the compound set dependence of feature annotation and the relatively small size of the available crystallographic reference sets. Accordingly, as a basis for compound selection, a ranked list in the case of IASF is expected to be less error prone than pre-defined threshold values.





**Figure 2.3: Outline of the IASF approach.** A flow diagram showing the different steps involved in fragment generation, interaction-based scoring and database searching. Detailed description is given and the major steps are summarized in the right boxes.



**Figure 2.4: Illustration of the flow chart of the IASF method.** In the upper left, the crystal structure of the inhibitor complex is shown (PDB ID 1SVG). The representation in the upper right focuses on a hydrogen bonding interaction involving carbonyl oxygen (atom number 18). The structure of the inhibitor is displayed in the right center. Carbonyl oxygen 18 is part of an ECFP4 feature computed for the ligand (consistently shown in magenta). Examples of ECFP4 features are shown at the middle bottom. On the basis of the calculated atom-based energy scores, the total score for the highlighted feature is obtained. If a database compound is found to contain this feature, it is assigned a score of -14.33.

### 2.2.1 Data set

The reference data set was assembled from the PDBbind database (Wang et al., 2004; Wang et al., 2005), an online accessible compilation of protein-ligand complexes extracted from the Protein Data Bank (PDB) (Berman et al., 2000). Reference ligands were selected from the “refined” subset of the PDBbind, which provides high-quality ligand structures selected for comparison of SBVS methods. Structural feature ensembles of ligands in complex crystal structures were generated using extended connectivity fingerprints with a bond radius of four (ECFP4) using the Scitegic Pipeline Pilot Student edition (v6.1.5.0), a program that serves as the connection of multiple pieces of software from different vendors, thus building a “pipeline”.

For the analysis, complex crystal structures of five target proteins with multiple ligands were chosen for which HTS data sets were also publicly available. The target systems include c-jun N-terminal kinase 3 (JNK), heat shock protein 90 (HSP), human immunodeficiency virus reverse transcriptase (HIV), protein kinase A (PKA), and thrombin (THR). For each protein, complexes with different inhibitors available in the PDB were selected, as summarized in Table 2.1. The 2D structures of the active reference compounds with the corresponding PDB codes are provided in Appendix B. Five screening datasets, one for each target protein, were obtained from the PubChem public database. These data sets ranged in size from ~60,000 (JNK) to ~217,000 (THR) tested compounds and contained between 62 (PKA) and 390 (HIV) biologically active hit compound.

**Table 2.1: Target proteins and screening data set.** 'Screening IDs' reports the PubChem bioassay identifiers and 'Total actives' and 'Total inactives' the number of hits and inactive compounds in each screening dataset, respectively. 'Inactive subset' gives the number of randomly selected inactive compounds used as background for VS calculations and 'PDB IDs' contains the list of protein-ligand complexes used.

Target protein	Screening IDs	Total actives	Total inactives	Inactive subset	PDB IDs	Docking template
PKA	524, 548	62	64,814	6,456	1Q8T, 1RE8, 1REJ, 1REK, 1YDS, 1YDT, 2ERZ, 1SVE, 1SVG, 1SVH	1XH8
THR	1046, 1215	223	216,693	21,929	1WAY, 1WBG, 2C8W, 2C8X, 2C8Y, 2C90, 2C93	1A4W
HIV	565, 651	390	63,969	6,066	1C0T, 1C0U, 1KLM, 1RT1, 1RT2, 1RTH, 1RTI, 1TKT, 1TKX, 1TKZ, 1TL1, 1TL3	1RTH
HSP	595	300	66,228	6,548	1UY7, 1UY8, 1UYC, 2VCI, 2VCJ	2BYI
JNK	746	366	59,422	5,883	1JNK, 1PMN, 1PMQ, 1PMU, 1PMV, 2B1P, 2EXC, 2O0U, 2O2U, 2OK1	2B1P

### 2.2.2 Atom-based scoring of protein-ligand interaction

Most flexible molecular docking programs utilize a scoring function as a measure of the free energy of binding to rank poses generated for a single ligand or ranking of different ligands relative to a receptor protein. It can also be used to analyze protein-ligand complexes to get detailed atom-based interaction information. The major challenge of scoring functions is among others, the failure to accurately predict solvation and entropy effects. But the effect of these two factors can be further simplified by taking individual terms of the scoring function that account for each atom of the ligand instead of taking the whole ligand. Therefore, to obtain scores of substructural features generated from X-ray crystal structure of protein-ligand complexes, individual components of the FlexX scoring function were used (Rarey et al., 1996; Rarey et al., 1997; Kramer et al., 1999). The FlexX scoring function is based on an empirical function, first reported by Böhm (Böhm, 1994), that estimates the free energy of binding. It is given as a sum of five different contributions:

$$\Delta G = \Delta G_0 + \Delta G_{rot} N_{rot} + \Delta G_{hb} \sum_{\text{neutral H-bonds}} f(\Delta R)f(\Delta \alpha) + \Delta G_{io} \sum_{\text{ionic interactions}} f(\Delta R)f(\Delta \alpha) + \Delta G_{aro} \sum_{\text{aromatic interactions}} f(\Delta R)f(\Delta \alpha) + \Delta G_{lipo} \sum_{\text{atom contacts}} f^*(\Delta R)$$

**Figure 2.5: The FlexX scoring function.**  $f(\Delta R, \Delta \alpha)$  is a scaling function penalizing deviations from the ideal geometry, and  $N_{rot}$  is the number of free rotatable bonds that are immobilized in the complex. The terms  $\Delta G_{hb}$ ,  $\Delta G_{io}$ ,  $\Delta G_{rot}$ ,  $\Delta G_{aro}$  and  $\Delta G_0$  are adjustable parameters accounting for the individual free energy contributions of the different interaction types.

The equation divides hydrogen-bond, salt bridge, and nonpolar interaction distances into a small number of discrete bins. Energy is assigned to each interaction based on which bin it occupies. Additional terms in the function are both entropic as well as enthalpic in nature, and consider the effect of buried surface area and the number of rotatable bonds in the ligand.

The scoring components are further divided in to 'match score' accounting for neutral hydrogen bonds, ionic and aromatic interactions as well as the 'contact score' accounting for lipophilic and van der Waals contacts. For example, the FlexX scoring function assigns an energy score of -4.7 kJ/mol to an ideal hydrogen bond, -8.7 kJ/mol to a strong ionic interaction, and -0.17 kJ/mol to a nonpolar van der Waals contact (for fragment scoring, energy units were omitted). Energy score components were selected that could be separated into per atom contributions in a meaningful manner.

Prior to scoring, the active site region for each ligand was defined by including residues within a 6.5 Å radius around each ligand atom. This radius

was chosen in order to ensure that longer range interactions were also taken into account.

### **2.2.3 Feature annotation**

For each ligand, the calculated energy score components were summed up for individual atoms using in-house python-based parsing script that parses the original FlexX raw energy scores. This enables the scoring of ECFP features generated from the reference set. ECFP4 features are by design in part overlapping and each feature must be independently annotated with the interaction information. For each ECFP4 feature, contributions of participating atoms were summed to generate the feature score. Features only occurring in a single ligand within an activity class were not included in the analysis and not scored. For features generated by multiple ligands, individual feature scores were averaged to obtain an activity class feature score.

### **2.2.4 Calculations**

For comparison of other methods with the new method, IASF, standard fingerprint search calculations were carried out using the ECFP4 features and MACCS keys applying the Tc coefficient as a weighing criterion. Each HTS data set was searched using the crystallographic inhibitors as reference molecules and the recall of active compounds among the by-Tc similarity top-ranked 500 screening set compounds was monitored. Enrichment factors over random selection were also calculated. As a fingerprint search strategy for multiple reference compounds, 1-NN nearest neighbor searching (Hert et al., 2004) was applied for both fingerprints. This means that for each screening set compound, the Tc similarity to each of the reference molecules is calculated and the highest value is selected as the final similarity score. In addition to similarity searching, flexible docking calculations were also carried out using the FlexX docking software applying the default parameter settings. In order to meet the computational expense of these calculations and enable a direct comparison with the LB methods, approximately 10% of the total number of inactive compounds were randomly selected from each HTS set (Table 2.1) and all hits were added. These subsets were used for all (i.e. IASF, fingerprint, docking, and substructure search) calculations. As docking template, the target protein with highest crystallographic resolution available in the PDB was used after removal of the bound ligand and bound crystal water molecules if any (Table 2.1). To rule out simple substructure matching, IASF calculations were also compared to substructure searching. For this purpose, maximum common substructures were extracted from the crystallographic ligands sharing identical or similar

scaffolds. These substructures were then used to search the screening sets and retrieve all molecules containing or sharing the substructures.

## **2.3 Results and Discussion**

### **2.3.1 The Interaction Annotated Structural Features (IASF) method**

Molecular fragments are typically used as binary (“present/absent”) descriptors in molecular similarity research, but have also been applied in weighted form, (Gillet and Willett, 1998; Durant et al., 2002; Jorgensen et al., 2006), for example, by calculating their frequency of occurrence in active compounds. The underlying idea of the IASF approach is to go beyond statistical analysis of fragment distributions and directly incorporate protein-ligand interaction information in fragment-based VS calculations. This enables the direct combination of 2D ligand similarity searching with 3D protein-ligand interaction information.

Incorporation of experimental protein-ligand interaction information into fragment matching is facilitated through the application of an energy function. The accuracy of energy-based scoring functions in SBVS is generally limited (Kitchen et al., 2004; Leach et al., 2006). In IASF calculations, partly overlapping structural features are scored focusing on selected interactions without the need to calculate a global free energy minimum, which represents the major limitation of scoring functions due to failure to accurately estimate the solvation, desolvation and entropy effects. Conceptually, IASF is best rationalized as a hybrid approach that combines 2D fragment matching with 3D protein-ligand interaction-based fragment weighting. Cumulative scoring on the basis of atom-based interaction energy values balances these contributions.

### **2.3.2 Analysis of the IASF method**

As reported in Table 2.1, between five (HSP) and 12 (HIV) different crystallographic ligands were used as reference molecules for feature generation and scoring. Comparable or larger numbers of complex structures with different ligands are available for a variety of target proteins in the PDB that could be subjected to IASF analysis. However, the choice of targets was largely determined by the availability of HTS data because it was intended to evaluate the approach on experimental screening data sets. Compared to artificially assembled compound benchmarking sets, screening data sets generally provide a more realistic and challenging basis for method comparisons because screening sets consist of experimentally confirmed active and inactive molecules. In

addition, screening hits are usually structurally diverse and chemically less complex than highly optimized active compounds that are often used for benchmarking and easier to distinguish from database compounds than screening hits.

In principle, IASF analysis can be carried out using any fragment ensembles and energy functions. Extended connectivity fingerprints were used here because they generate feature sets for individual molecules that are more specific than for example, pre-defined fragment dictionaries. Furthermore, the FlexX scoring function was chosen for two reasons. FlexX energy component values can be easily parsed into individual atomic contributions and, in addition, the FlexX scoring function emphasizes both uncharged and charged hydrogen bond interactions. The latter aspect was considered important for feature annotation because we intended to primarily capture molecular recognition and specificity determinants that can be directly assigned to ligand fragments. By contrast, 3D scoring shape complementarity between ligand and target is difficult on the basis of 2D substructures and is here only partly and indirectly accounted for through the inclusion of non-polar van der Waals contacts (and steric overlap penalties).

The feature and score distributions for the five compound classes are reported in Table 2.2 below. Between 74 (HSP) and 110 (PKA) annotated ECFP4 features were generated per ligand set. These features were used for IASF similarity searching, as discussed below.

**Table 2.2: Feature and score distributions.** 'Annotated Features' gives the total number of accepted ECFP4 features per compound reference set (i.e. features occurring in at least two reference molecules) and 'Features per Ligand' the average number of generated features per crystallographic reference molecule. 'Atoms min' and 'max' give the minimum and maximum number of atoms per feature, respectively. 'Score min', 'max', and 'avg' report the minimum (best), maximum, and average scores per reference set, respectively.

Target protein	Annotated features	Features per ligand	Atoms min	Atoms max	Score min	Score max	Score avg.
PKA	110	10.0	1	10	-18.29	-0.32	-6.91
THR	79	11.2	1	11	-11.58	-0.13	-3.65
HIV	93	7.8	1	10	-15.33	-0.33	-5.71
HSP	74	14.8	1	9	-14.13	-0.14	-4.38
JNK	105	10.5	1	10	-23.97	0.00	-5.50

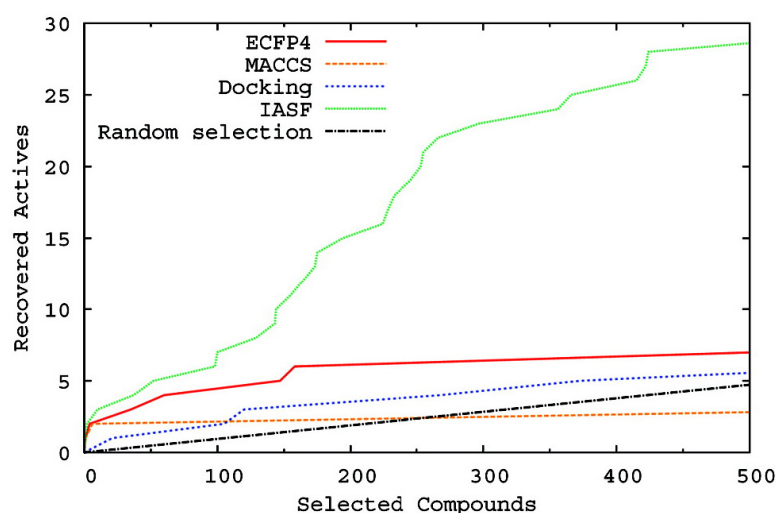
The average number of features per ligand ranged from 7.8 (HIV) to 14.8 (HSP), and individual features contained between one and 11 non-hydrogen atoms. Feature scores were of comparable magnitude for the different ligand sets and in each case, features were found with scores close or equal to

zero (which means that they were essentially not involved in interactions accounted for by the score components). This gives a clear map of interacting and non-interacting fragments which can be used to directly apply the score annotated ligand fragments as fingerprints for use in similarity searching.

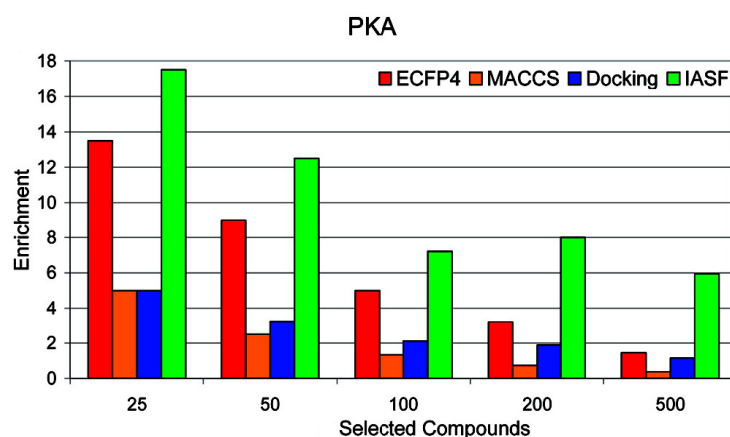
### 2.3.3 Evaluation of the performance of IASF

Following feature generation and annotation, IASF was applied to mine compounds from five HTS data sets. The performance of IASF was then compared with the selected reference methods, ECFP4 and MACCS fingerprint similarity searching and FlexX molecular docking. The results are summarized in Figures 2.6 - 2.10 below.

(a)



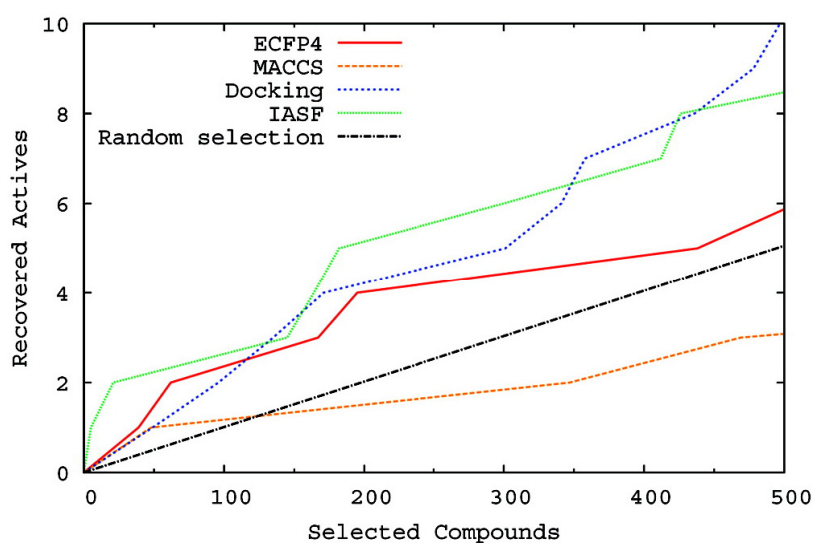
(b)



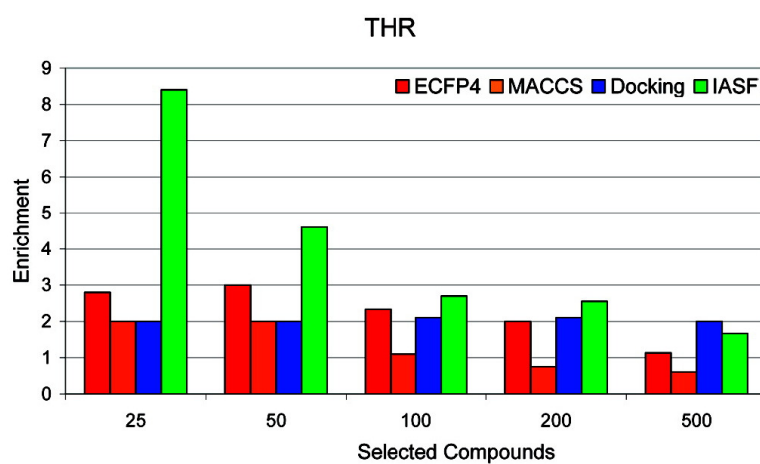
**Figure 2.6: Virtual screenings trials for PKA.** Two graphs are shown that report recall curves (a) and enrichment factors (b) for search calculations using IASF and the selected reference methods. The color coding of the respective graphs is described in the legend.



(a)

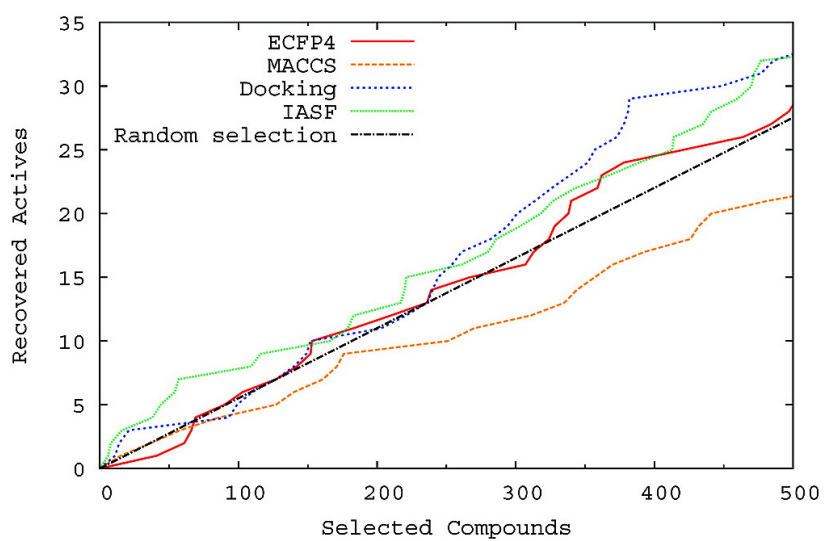


(b)

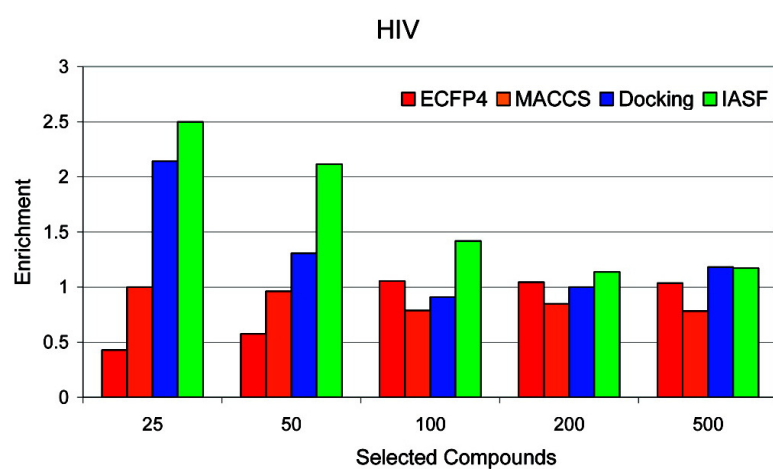


**Figure 2.7: Virtual screenings trials for THR.** Two graphs are shown that report recall curves (a) and enrichment factors (b) for search calculations using IASF and the selected reference methods. The color coding of the respective graphs is described in the legend.

(a)

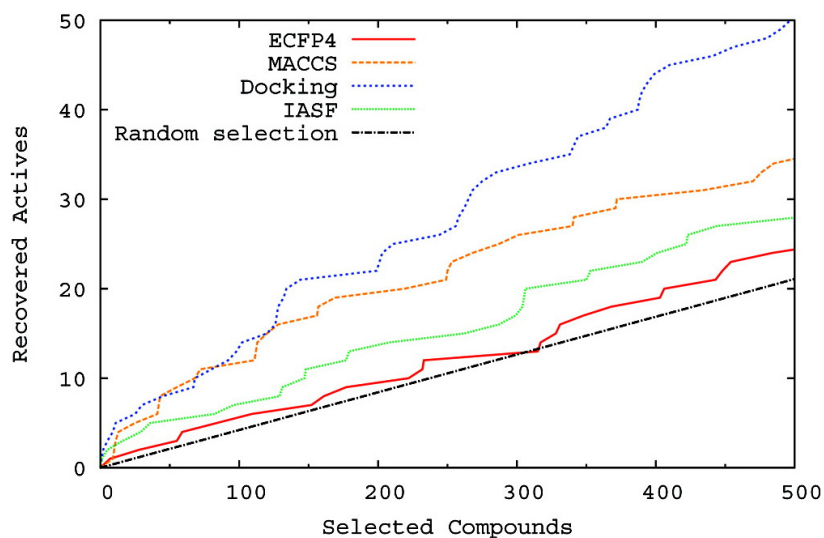


(b)

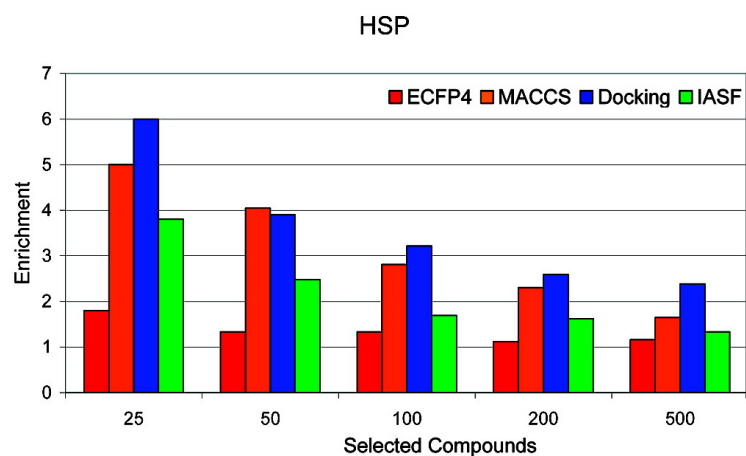


**Figure 2.8: Virtual screenings trials for HIV.** Two graphs are shown that report recall curves (a) and enrichment factors (b) for search calculations using IASF and the selected reference methods. The color coding of the respective graphs is described in the legend.

(a)

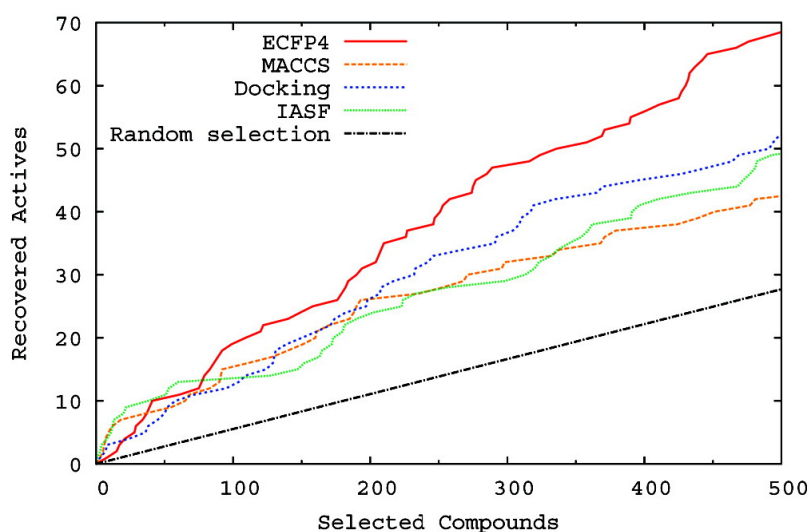


(b)

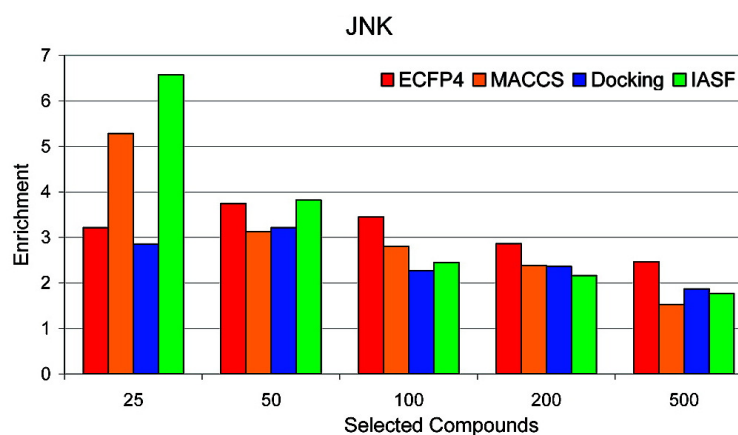


**Figure 2.9: Virtual screenings trials for HSP.** Two graphs are shown that report recall curves (a) and enrichment factors (b) for search calculations using IASF and the selected reference methods. The color coding of the respective graphs is described in the legend.

(a)



(b)



**Figure 2.10: Virtual screenings trials for JNK.** Two graphs are shown that report recall curves (a) and enrichment factors (b) for search calculations using IASF and the selected reference methods. The color coding of the respective graphs is described in the legend.

As expected, these screening sets provided challenging test cases for all methodologies. Active compounds were retrieved in essentially all calculations but their frequency was only a two- to three-fold enrichment over random selection. IASF produced overall highest compound recall and enrichment factors in three of the five cases, PKA, THR, and HIV. In the case of PKA (Figure 2.6a and b), IASF clearly dominated the calculations. On THR (Figure 2.7a and b), IASF outperformed the reference methods for small selection sets.

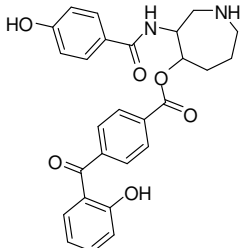
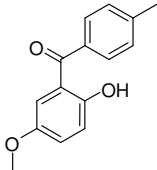
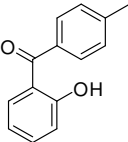
For HIV (Figure 2.8a and b), only IASF and docking calculations produced meaningful recall and enrichment factors. IASF performance was lowest in the case of HSP (Figure 2.9a and b) where only five crystallographic ligands were utilized for fragment generation and annotation. However, IASF was consistently superior to ECFP4 in this case. Here docking produced highest recall followed by MACCS keys. For JNK (Figure 2.10a and b), IASF produced highest recall of hits and enrichment factors for the first 50 screening set compounds. For larger sets, ECFP4 searching produced higher recall but the enrichment factors were comparable. Overall, IASF performed best on these five test cases, in particular, for small compound selection set sizes. Importantly, the performance of IASF was generally superior to ECFP4 searching, although ECFP4 generated a total of approximately 300 to 500 features for each of the five reference sets (compared to between 74 and 110 annotated IASF features). In four of the five cases, IASF was also superior to docking calculations. These findings demonstrate the gain in target-specific information achieved by 3D interaction annotation of ECFP4 features. IASF calculations displayed a notable tendency to enrich active compounds in relatively small database selection sets. This behavior suggests that annotation with interaction information renders search calculation specific for subsets of active compounds that engage in similar interactions. Thus, as intended, IASF calculations focus search calculations on selected structural features (and thereby depart from general structural feature matching).

#### **2.3.4 Comparison of IASF and substructure searching**

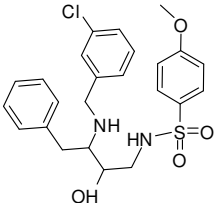
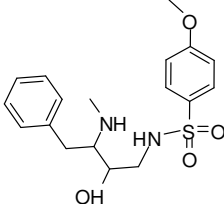
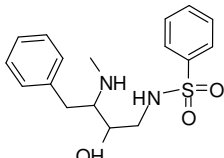
For comparison and to rule out any random substructure matching in the IASF method, substructure search calculations were performed using the largest common substructures shared by subsets of crystallographic reference ligands containing similar scaffolds. The results indicated that the substructure search calculations essentially failed to retrieve active compounds from the screening sets. The results are summarized in Table 2.3.

**Table 2.3: Summary of the maximum common substructure searching results.** Substructures and number of active and inactive compounds retrieved for (a) PKA, (b) THR, (c) HIV, (d) HSP and (e) JNK. 'No. of ligands' reports the number of crystallographic ligands from which the displayed maximum common substructure was derived. 'No. of actives retrieved' and 'No. of inactives retrieved' give the number of hits and inactive compounds retrieved from each screening set, respectively. In contrast to virtual screening calculations, the entire screening sets were used for the substructure searching.

(a)

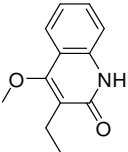
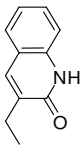
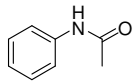
No. of ligands	Max. Common substructure	No. of actives retrieved	No. of inactives retrieved
2		0	0
5		0	1
5		0	5

(b)

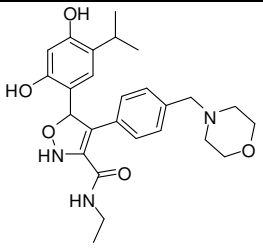
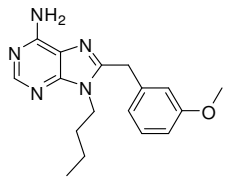
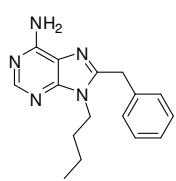
No. of ligands	Max. Common Substructure	No. of actives retrieved	No. of inactives retrieved
2		0	0
3		0	0
4		0	0

**Table 2.3: (continued) Summary of the maximum common substructure searching results.** Substructures and number of active and inactive compounds retrieved for (a) PKA, (b) THR, (c) HIV, (d) HSP and (e) JNK. 'No. of ligands' reports the number of crystallographic ligands from which the displayed maximum common substructure was derived. 'No. of actives retrieved' and 'No. of inactives retrieved' give the number of hits and inactive compounds retrieved from each screening set, respectively. In contrast to virtual screening calculations, the entire screening sets were used for the substructure searching.

(c)

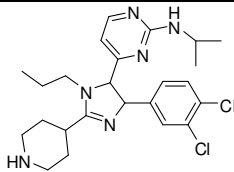
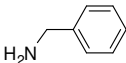
No. of ligands	Max. Common Substructure	No. of actives retrieved	No. of inactives retrieved
3		0	2
5		0	210
6		56	13,730

(d)

No. of ligands	Max. Common Substructure	No. of actives retrieved	No. of inactives retrieved
1		0	0
2		0	0
3		0	0

**Table 2.3:** (continued) **Summary of the maximum common substructure searching results.** Substructures and number of active and inactive compounds retrieved for (a) PKA, (b) THR, (c) HIV, (d) HSP and (e) JNK. 'No. of ligands' reports the number of crystallographic ligands from which the displayed maximum common substructure was derived. 'No. of actives retrieved' and 'No. of inactives retrieved' give the number of hits and inactive compounds retrieved from each screening set, respectively. In contrast to virtual screening calculations, the entire screening sets were used for the substructure searching.

(e)

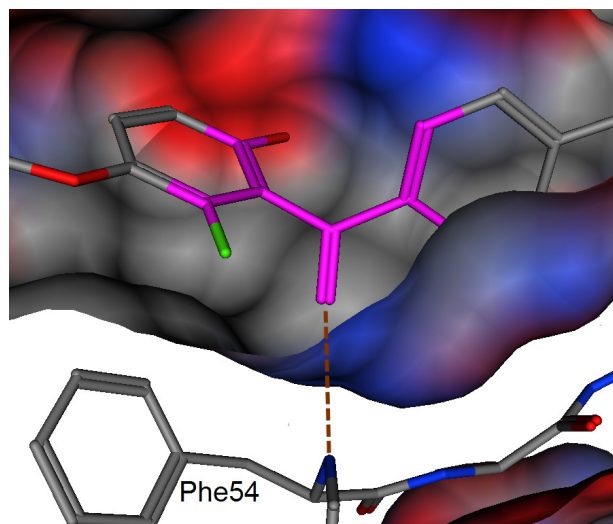
No. of ligands	Max. Common Substructure	No. of actives retrieved	No. of inactives retrieved
2		0	0
5		107	15,203

As shown in Table 2.3 above, in 12 of the 14 substructure search calculations performed, no hits were identified. In two calculations, on HIV and JNK, 56 and 107 active compounds were retrieved together with 13,730 and 15,203 inactive compounds, respectively. Thus, the maximal common substructure searching did not provide meaningful search results. These findings demonstrate that the IASF calculations are not comparable to simple substructure searching.

### 2.3.5 Analysis of annotated features

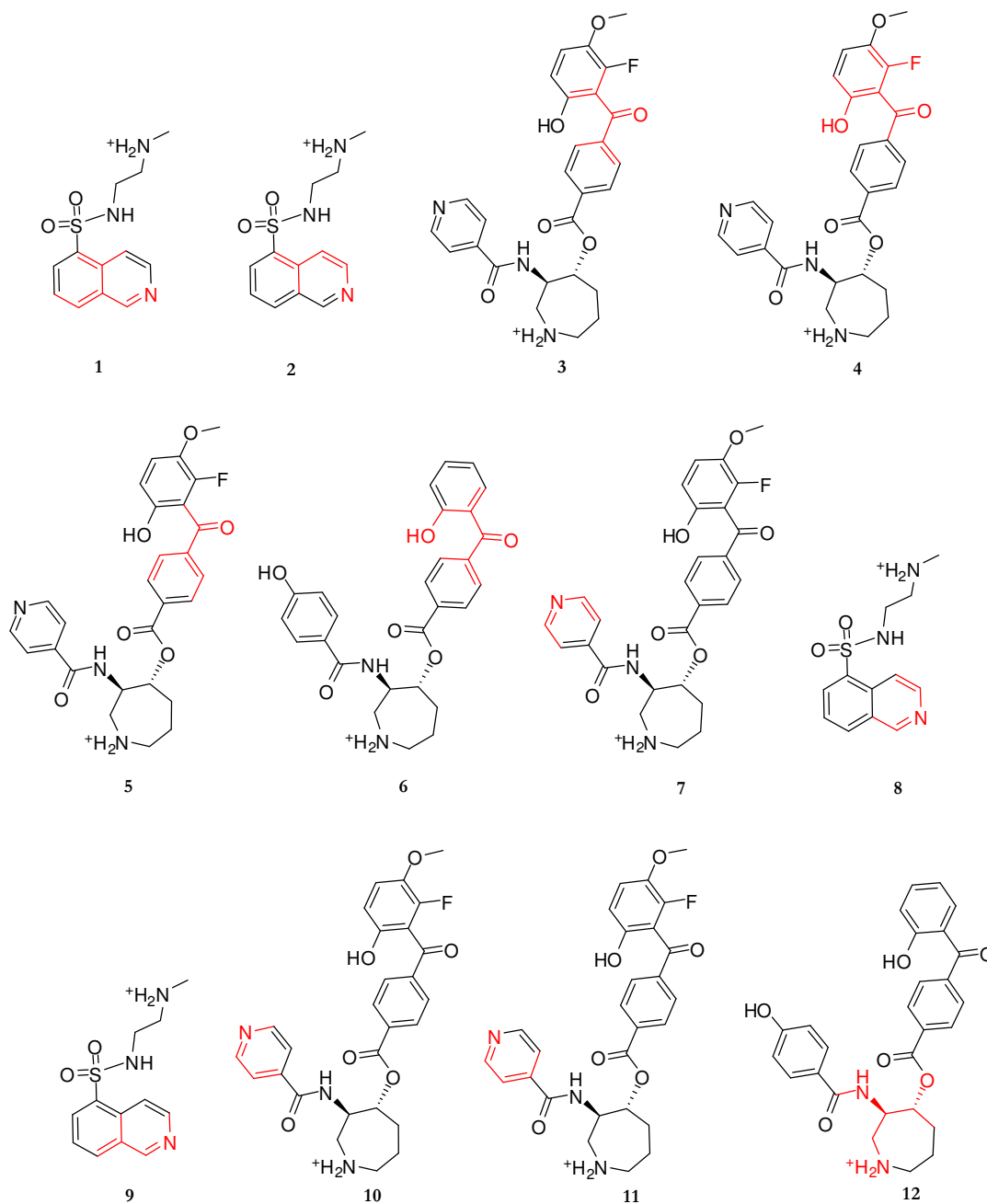
Selected substructural features, generated by the IASF method, were further investigated to get a detailed understanding of their nature and the 3D interactions they form. For example, the substructure highlighted in Figure 2.7 below obtained the third best score among all PKA features. It is part of the benzophenone motif that is a signature for this series of PKA inhibitors (Breitenlechner et al., 2004).





**Figure 2.7: Key interaction between a ligand substructure and PKA.** A close-up view of the carbonyl oxygen of the benzophenone moiety forming a hydrogen bond with the amide of the protein residue Phe54 is shown. The receptor is presented in surface with atom-type coloring.

In the crystal structure, the carbonyl oxygen of the benzophenone moiety forms a hydrogen bond to the backbone amide NH of PKA residue Phe54. Figure 2.8 below shows the top 12 PKA features in the context of the inhibitor structures they were derived from. All of these features contain hetero atoms that function as hydrogen bond donor or acceptors, consistent with the choice of the FlexX score components and the proposed expectation.



**Figure 2.8: Top scoring 12 PKA substructural features.** The features are highlighted in red in the context of their inhibitor structures they were derived from.

For example, the top two features contain major portions of the isoquinoline moiety including the nitrogen that forms a hydrogen bond to the adenine ring of ATP, the kinase cofactor. Feature annotation also discriminates between more and less conserved interactions. For example, if a heteroatom within a feature forms a hydrogen bond in only one of several inhibitors it occurs in, as also observed for PKA inhibitors of the benzophenone and azepine series depicted in Figure 2.8, the average absolute value of the score of this feature is low for the reference set. Thus, substructures involved in

interactions that are conserved among reference set inhibitors are most highly weighted.

## **2.4 Summary**

In this study, a new 2D/3D hybrid methodology was introduced that adds protein-ligand interaction information to sets of substructures derived from active compounds. Interaction information is extracted from crystallographic data through the application of an energy function. Annotated substructures are then used to search databases for retrieval active compounds. Database compounds are assigned cumulative scores based on substructures they share with active reference compounds and the associated energy scores. The methodology is ligand-centric because it relies on mapping of interaction-weighted substructures. These substructures represent the most important activity specific features. In benchmark calculations on different HTS data sets, the hybrid approach mostly performed better than 2D (fingerprint) and 3D (docking) calculations. These findings suggest that substructure and interaction knowledge is highly complementary in nature and that there is considerable gain in structure-activity relationship information when these 2D structures and 3D interaction components are combined.



# Chapter 3

## Structural interpretation of activity cliffs revealed by systematic analysis of SARs in analog series

In the previous chapter, a new methodology was introduced that utilizes 3D ligand-target interaction information to enhance the performance of 2D-based fingerprint similarity searching. The data showed that high scoring substructures extracted from active ligands represent the most important activity specific features. In benchmark calculations on different HTS data sets, the hybrid approach mostly performed better than 2D (fingerprint) and 3D (docking) calculations. The findings suggested the existence of high complementarity between ligand substructure and 3D interaction knowledge where considerable gain in structure-activity relationship information is observed when 2D structures and 3D interaction components are combined. This chapter focuses on a continued application of protein-ligand interaction information but this time for studying activity cliffs revealed by systematic analysis of SARs in analog series (Sisay et al., 2009a). Study design, methodology and results are discussed in detailed.

### 3.1 Introduction

Understanding the structure-activity relationships (SAR) of a set of compounds with measured biological activity plays a key role in VS. In hit-to-lead and lead optimization projects, active compounds are subjected to chemical modification and series of analogs are generated from which SAR information is

extracted. In analog design and exploration, one typically attempts to identify substitution sites where substituent (R-group) variations lead to improved potency (and other desired compound physicochemical characteristics) and aims to identify SAR patterns that can be rationalized and ultimately used to predict highly potent compounds (Peltason and Bajorath, 2009). However, the derivation of SAR rules and guidelines for optimization is often severely compromised by abrupt changes in the biological response to minor chemical modifications of active compounds. This situation is typically attributed to the presence of an activity cliff (Maggiora, 2006) or a region of SAR discontinuity (Peltason and Bajorath, 2009). Both of these terms are derived from the intuitive concept of an activity landscape (Maggiora, 2006; Peltason and Bajorath, 2009; Bajorath et al., 2009) that describes activity responses to positional changes in biologically relevant chemical space. At activity cliffs, small changes in structure, corresponding to small steps in chemical space, lead to significant changes in the activity or potency hypersurface. Multiple activity cliffs can be present in the activity landscape shaped by a compound series (Schneider, G. and Schneider, P., 2004; Guha and van Drie, 2008; Peltason and Bajorath, 2009; Bajorath et al., 2009, Medina-Franco et al., 2009) and each of these cliffs gives rise to local SAR discontinuity observed for a compound subset. Accordingly, the terms activity cliff and SAR discontinuity are conceptually linked and can essentially be used interchangeably. The magnitude of activity cliffs is generally influenced by the chosen chemical reference space and molecular representation. The same applies to the location of different activity islands, i.e. small regions in chemical space that are enriched with compounds sharing a specific biological activity.

SAR analysis functions (Bajorath et al., 2009) such as the SAR Index (SARI) (Peltason and Bajorath, 2007a) have been designed to quantitatively account for SAR continuity and discontinuity within compound data sets. In principle, a compound set is characterized by low SAR discontinuity if it consists of moderately similar or even structurally diverse compounds having only relatively small differences in potency and, in contrast, by high discontinuity if it contains very similar compounds with dramatic potency differences.

A systematic study of such SAR patterns within analog series is generally complicated because SARs are typically heterogeneous in nature (Bajorath et al., 2009), i.e. they consist of multiple components and often combine continuous and discontinuous elements. In order to study local SAR features in analog series, a data structure termed combinatorial analog graph (CAG) was recently developed by Peltason *et al.* (Peltason et al., 2009) that systematically divides analogs into subsets of compounds that only differ at defined substitution sites

and, in addition, incorporates the SARI scoring scheme to identify substitution patterns that are responsible for SAR discontinuity.

CAGs and SAR analysis functions exclusively utilize similarity and potency information of active compounds as input, i.e. information encoded at the ligand level. However, SAR information contained in analog series is naturally to a large extent determined by underlying receptor-ligand interactions (Peltason and Bajorath, 2009), which are the focal point of SB ligand design efforts (Jorgensen, 2004). Given this close link between receptor-ligand interactions and compound activity or potency, one might perhaps expect that relationships between molecular structure and biological activity observed at the ligand level could be easily reconciled on the basis of 3D protein-ligand interaction information. However, it has been demonstrated that relationships between 2D similarity, 3D (binding mode) similarity, and potency of active compounds are often highly complex and difficult to predict (Peltason and Bajorath, 2007b). This emphasizes the fact that interactions seen in receptor-ligand complexes are only one of several factors that determine or influence SARs. Nevertheless, the presence of strong SAR discontinuity is often thought to be a direct consequence of compromised receptor-ligand interactions.

In this study, in order to further understand the relationship between SAR discontinuity and protein-ligand interactions, analysis of analog sets for which 3D structural information was available was carried out. This made it possible to interpret SAR information extracted from active compounds projected into chemical reference spaces at the target structural level. For the analysis, series of analogs directed against different targets were systematically analyzed in CAGs in order to identify substitution patterns that were directly responsible for SAR discontinuity within these series. Key substitution sites were then analyzed on the basis of crystallographic data to map activity cliffs and rationalize SAR discontinuity within the framework of specific receptor-ligand interactions.

## 3.2 Methodology

### 3.2.1 Compound analog series and X-ray structures

For the analysis, five series of structural analogs of four target enzymes for which one of the analogs was available in a complex X-ray crystal structure was selected (Table 3.1). The analog series included inhibitors of carbonic anhydrase II (PDB code 2HOC), Tie-2 kinase (2P4I), factor Xa (2BMG and 2G00), and thrombin (1SL3). Active compounds and corresponding potency data were taken from the BindingDB public database (Liu et al., 2007), with the exception

of the thrombin inhibitor series that was taken from the literature (Young et al., 2004).

**Table 3.1: Summary of study data set.** In the column 'PDB code', the PDB code of the crystallographic enzyme-inhibitor complex is given, and the column 'PDB ligand id' reports the unique PDB identifier of the X-ray ligand. '# of Cpds' denotes the number of compounds. 'CA II': carbonic anhydrase II.

Target protein	PDB code	PDB ligand id	# of Cpds	Potency range [nM]
CA II	2HOC	1CN	6	0.3 - 9
Tie-2 Kinase	2P4I	MR9	8	1 - 399
Factor Xa(1)	2BMG	I1H	20	13 - 1053
Factor Xa(2)	2G00	4QC	13	0.18 - 88
Thrombin	1SL3	170	13	0.0015 – 940

### 3.2.2 Analysis of 3D protein-ligand interactions

For the analysis of the interaction of each analog with the corresponding enzyme, the reference protein-ligand complex was imported into the MOE graphical environment and repeated chains or any unwanted structures were removed. All water molecules, except those involved in ligand interaction were also removed. The active site was defined by taking all residues within 6.5 Å measured from all atoms of the bound X-ray inhibitor. Enzyme-inhibitor interactions in X-ray structures were analyzed with MOE by applying an additional 3D protein-ligand interaction analysis module. Following that each R-group was attached to the reference compound and the resulting local interaction pattern including van der Waals clashes were thoroughly investigated. This was done by using the MOE-3D interaction analysis module and thorough manual inspection of the local active site interaction properties of the protein. In addition literature information on the physicochemical characteristics of each analog and the active site properties of the protein was taken into consideration. For the structural correlation analysis presented herein, SAR data was obtained from enzyme-based inhibition assays and not other assay formats.

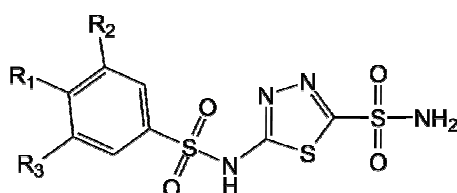
### 3.2.3 R-group decomposition

For each analog series, the maximum common subgraph (MCS) shared by all analogs in the series was determined. The MCS was then utilized as the invariant core structure for R-group decomposition to determine (and consistently



number) corresponding substitution sites in analogs and assign sets of substituents (functional groups) to these sites. For the analog series studied herein, all substitution sites were unambiguously assigned. MCS calculation and R-group decomposition were performed using the Pipeline Pilot software. Table 3.2 reports core structure, substitution sites, and R-groups for carbonic anhydrase II series (the CAG generation was performed by Dr. L. Peltason). The SAR tables including core structures, substitution sites, and R-groups for the rest of the four series discussed in this chapter can be found in Appendix C.

**Table 3.2: Carbonic anhydrase II inhibitor series core structure and corresponding R-groups after decomposition.** Molecular frameworks and consistently numbered R-groups are presented. For individual compounds, substituents and potency values are reported. Attachment atoms are labeled with 'Z'. Compounds from the BindingDB are identified by their unique BindingDB monomer id. For the reference X-ray ligand, the PDB ligand identifier is given in parentheses.



BindingDB monomer id	Potency [nM]	R1	R2	R3
10870	2	Z-NH <sub>2</sub>		
10886	9			
11621	0.8	Z-NH <sub>2</sub>	Z-F	
11622	0.6	Z-NH <sub>2</sub>		Z-Cl
11625 (1CN)	0.3	Z-NH <sub>2</sub>	Z-F	Z-Cl
11628	0.5	Z-NH <sub>2</sub>	Z-Cl	Z-Cl

### 3.2.4 Organization and analysis of analog series

In order to quantify contributions of substitution sites to SAR discontinuity, each analog series was systematically divided into subsets of compounds that only differed at a specific substitution site or combinations of up to three sites. For the resulting compound subsets, the SARI discontinuity score (Peltason and Bajorath, 2007a) was calculated and CAGs (Peltason et al., 2009) were used to organize analog subsets and visualize contributions of substitution patterns to SAR discontinuity.

### 3.2.5 SARI discontinuity scores

The SARI discontinuity score (Peltason and Bajorath, 2007a) calculates pairwise potency differences between analogs and averages them to obtain a measure of the magnitude of potency differences among similar compounds. The pairwise potency differences are scaled by the similarity value of the corresponding compound pair in order to emphasize potency differences between highly similar compound pairs:

$$\text{disc} = \text{mean}_{\{(i,j)|i \neq j\}} \left( |P_i - P_j| \times \text{sim}(i, j) \right) \quad (3.1)$$

Here,  $P_i$  and  $P_j$  denote the potency values of compounds  $i$  and  $j$  and  $\text{sim}(i, j)$  denotes their similarity, calculated as the Tc (Willett et al., 1998) for MACCS fingerprint representations. SARI scoring has been found to be rather stable for compound reference sets of varying size and different molecular representations including structural keys and topological fingerprints (Peltason and Bajorath, 2009). For analyzing analog series, the application of a similarity threshold value for calculating the discontinuity score is not required because analogs have by definition highly similar structures. Therefore, the chosen molecular representation is also not critical in this case as long as it correctly counts for the high similarity of the compared compounds (which is certainly the case for structural keys).

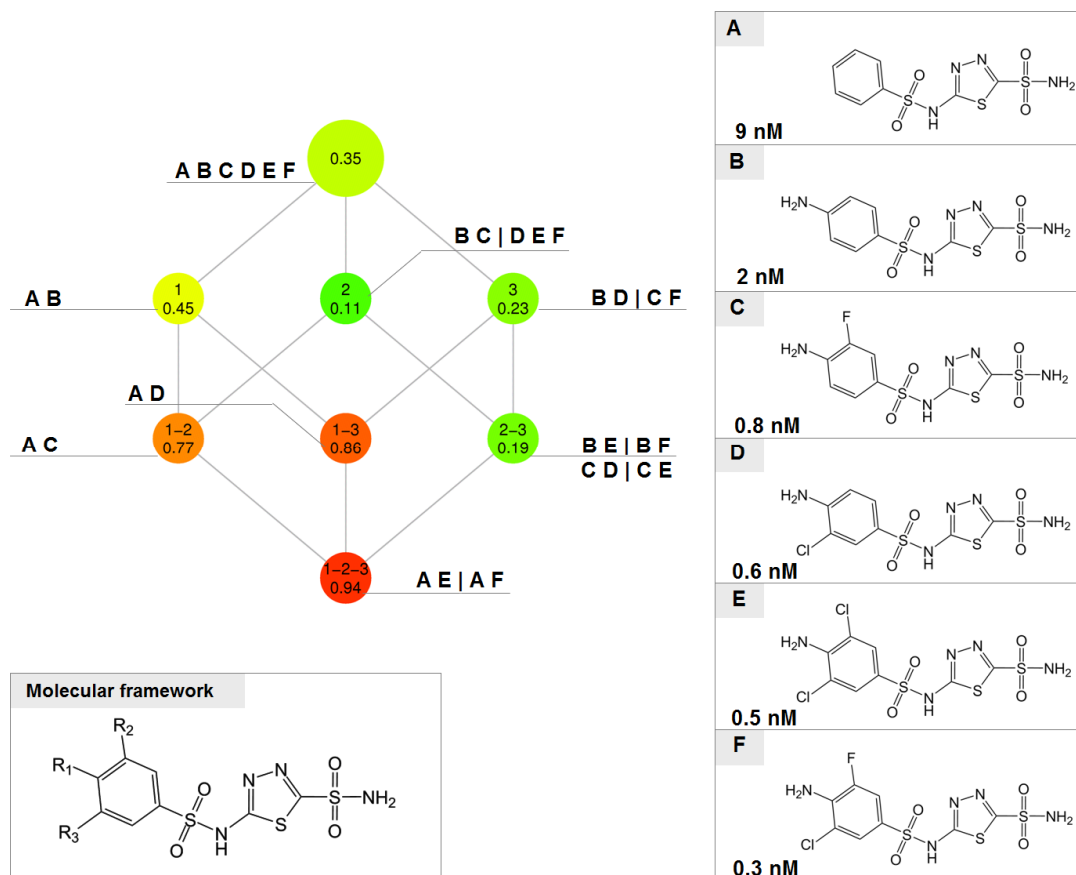
The discontinuity score is calculated for an entire analog series and all of its subsets. Scores are normalized and mapped to the value range [0, 1] as described previously (Peltason and Bajorath, 2007a). Score values close to 1 account for a strongly discontinuous SAR, whereas low values close to 0 reflect a low degree of SAR discontinuity. Here, scores of all compound subsets within an analog series serve as reference for normalization of the series. Thus, score distributions are characteristic of a given analog series and account for individual potency distributions.

## 3.3 Results and Discussion

### 3.3.1 The CAG formulation

The CAG representation organizes analog series as compound subsets having modifications exclusively at defined sites or site combinations and accounts for SAR discontinuity within subsets that can be directly attributed to modifications at the given substitution sites. Figure 3.1 shows a simple CAG representation for a small set of six carbonic anhydrase II inhibitors. Here individual analogs

are shown and the subsets they form at different nodes are reported, illustrating the fact that analogs usually participate in different subsets, given the distribution of substituents.



**Figure 3.1: CAGs and SAR discontinuity analysis:** Combinatorial analog graph for the inhibitor series of carbonic anhydrase II is shown. Nodes in the CAG are labeled with substitution sites and site combinations and are color-coded according to the degree of SAR discontinuity within the corresponding compound subsets. All inhibitor structures are shown and the nodes are additionally given the corresponding inhibitor label in order to illustrate the composition of overlapping compound subsets corresponding to the nodes. If more than one subset is available for a node (e.g. at nodes 2, 3, 2-3 and 1-2-3), individual subsets are separated by a vertical line.

As exemplified in Figure 3.1, a node might correspond to several subsets that consist of compounds that differ only at the given sites but are distinguished at another site. Discontinuity scores for these subsets are calculated independently and averaged. In the CAG, the top (root) node represents the entire analog series and each non-root node represents a subset of compounds with different substitutions at the specified sites. Node labels identify the substitution sites and report discontinuity scores for the compound subset representing each site combination. For example, “1” and “1-2” means that compound subsets only differ at site 1 or sites 1 and 2, respectively, but are

otherwise identical. Nodes are arranged in layers according to the number of substitution sites that are considered and color-coded according to discontinuity scores using a spectrum from green (score 0, i.e. no SAR discontinuity) over yellow to red (score 1, i.e. maximal discontinuity). Edges are drawn from a node to all other nodes in the next layer whose substitution site combination includes all of the sites represented by the originating node. Substitution site combinations for which no compounds are available (i.e. non-explored combinations) are shown as small white nodes. Combinations of up to three sites are systematically accounted for and the complexity of a CAG increases with the number of individual sites. For example, for three substitution sites, one bottom node with a three-site combination is obtained but for four sites, there are four bottom nodes.

### 3.3.2 Patterns of SAR discontinuity and mapping of activity cliffs

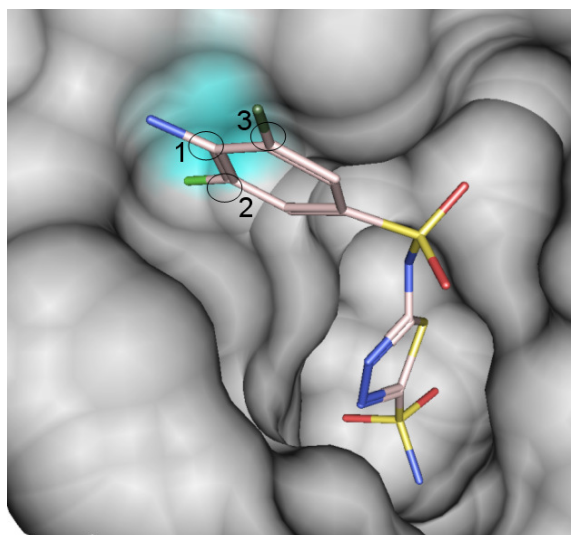
For each of the five analog series, CAG representations were generated and in each case, substitution patterns were identified that introduced SAR discontinuity. It was found that discontinuity patterns substantially varied among these analog series. In the following sections, the CAG representations of the different series are discussed and substitution sites that introduce SAR discontinuity are evaluated in the context of protein-ligand interactions.

**Carbonic anhydrase II:** The six analog carbonic anhydrase inhibitors including the X-ray ligand range in potency from 0.3 nM - 9 nM and differ at three substitution sites. The CAG in Figure 3.1 clearly shows that individual modifications at single substitution sites do not lead to significant potency differences (green nodes 1, 2, and 3 at the first level). By contrast, combinations of modifications at sites 1-2 and 1-3 (but not 2-3) result in potency differences of more than one order of magnitude, and largest SAR discontinuity is observed for simultaneous modifications at sites 1-2-3 (red node at the bottom). Thus, given the SAR information that is extracted from overlapping compound subsets, this series would be suggested to contain substitutions that depend on each other and act in concert.

The X-ray structure of the enzyme-inhibitor complex, shown in Figure 3.2, indicated that substitutions at site 1 point outside the active site and are thus not expected to have significant interaction with the protein. Furthermore, the complex reveals information that could not be deduced from SAR analysis of this analog series.

Importantly, for substitutions at sites 2 and 3 at the phenyl ring that is freely rotatable, only one small hydrophobic binding pocket exists which is

formed by the residues Val135, Leu198, Pro202 and Leu204. In this compound series, filling this pocket represents an activity cliff for strong inhibition. However, this can be accomplished by a halogen substituent at either site 2 or 3. Thus, substitutions at these sites do not act in a concerted manner, as suggested by analyzing the series, but in an alternative way.



**Figure 3.2: Complex crystal structure of carbonic anhydrase II with inhibitor 11625 (numbered according to Table 3.1; PDB code 2HOC).** In this structural representation, the active site of the enzyme is depicted as a solid surface. A selected hydrophobic pocket is shown with cyan surface coloring. The inhibitor is presented in stick using the atom color codes: light-rose for carbon; red for oxygen; blue for nitrogen; yellow for sulfur; green for fluorine and dark-green for chlorine. Substitution sites in the inhibitor are circled and labeled accordingly.

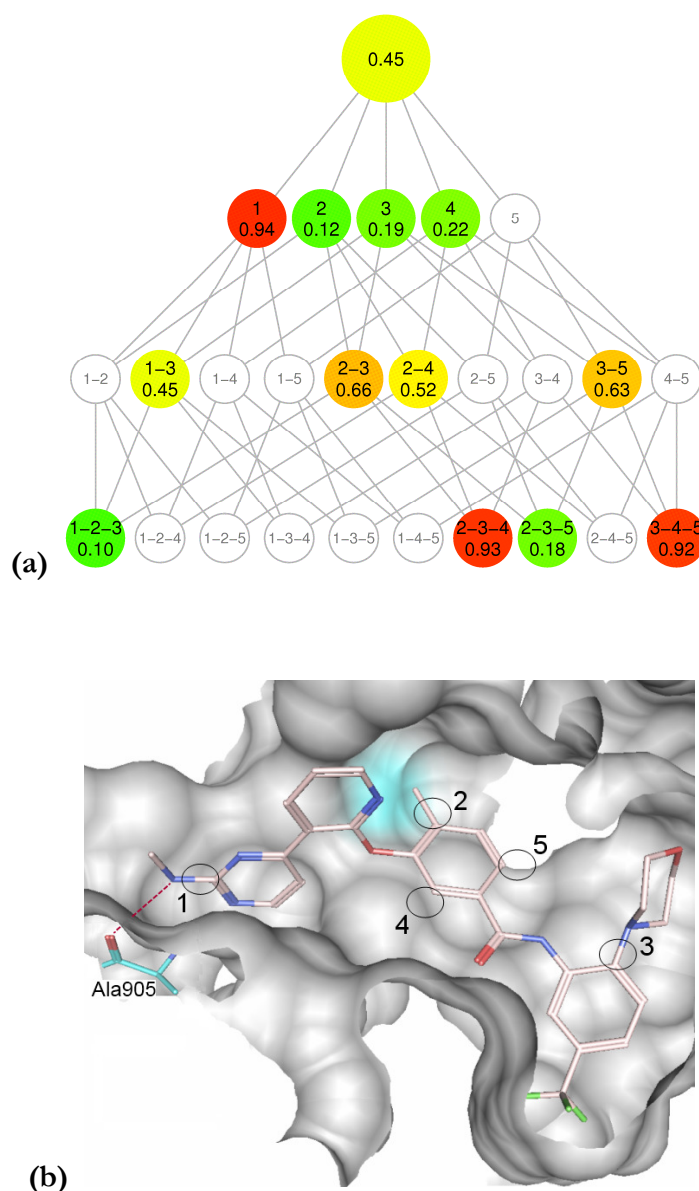
This information could not be deduced from the CAG because halogen-substituted compounds at site 2, 3, or both sites have comparable potencies, consistent with the 3D protein-ligand interaction picture, and influence SAR discontinuity in similar ways.

**Tie-2 kinase:** This series consists of eight analogs that are ATP-site directed inhibitors. The analogs differ at five substitution sites, and fall into the potency range of 1 nM - 399 nM. As illustrated by the CAG in Figure 3.3a, many potential combinations of substitution sites are currently unexplored (shown as “empty” nodes).

However, substitution site 1 emerges as a prominent hotspot for individual modifications. For example, the addition of a methylamine at site 1 increases potency by up to two orders of magnitude (Table 3.1). This can be easily explained considering the structure and the protein-ligand complex shown in Figure 3.3b. The substituent at site 1 forms a strong hydrogen bond to the backbone carbonyl oxygen of Ala905 at the bottom of the pocket. The

interaction with this residue that is conserved in many kinases represents a prominent activity cliff for ATP site-directed inhibition.

Furthermore, modifications at sites 2-3 and 2-4 introduce moderate SAR discontinuity but combinations of the same modifications at sites 2-3-4 yield a high degree of discontinuity. Thus, modifications at sites 2-3 and 2-4 have additive effects. Modifications of sites 2 and 4 include the presence or absence of a methyl group and only have a minor SAR effect.

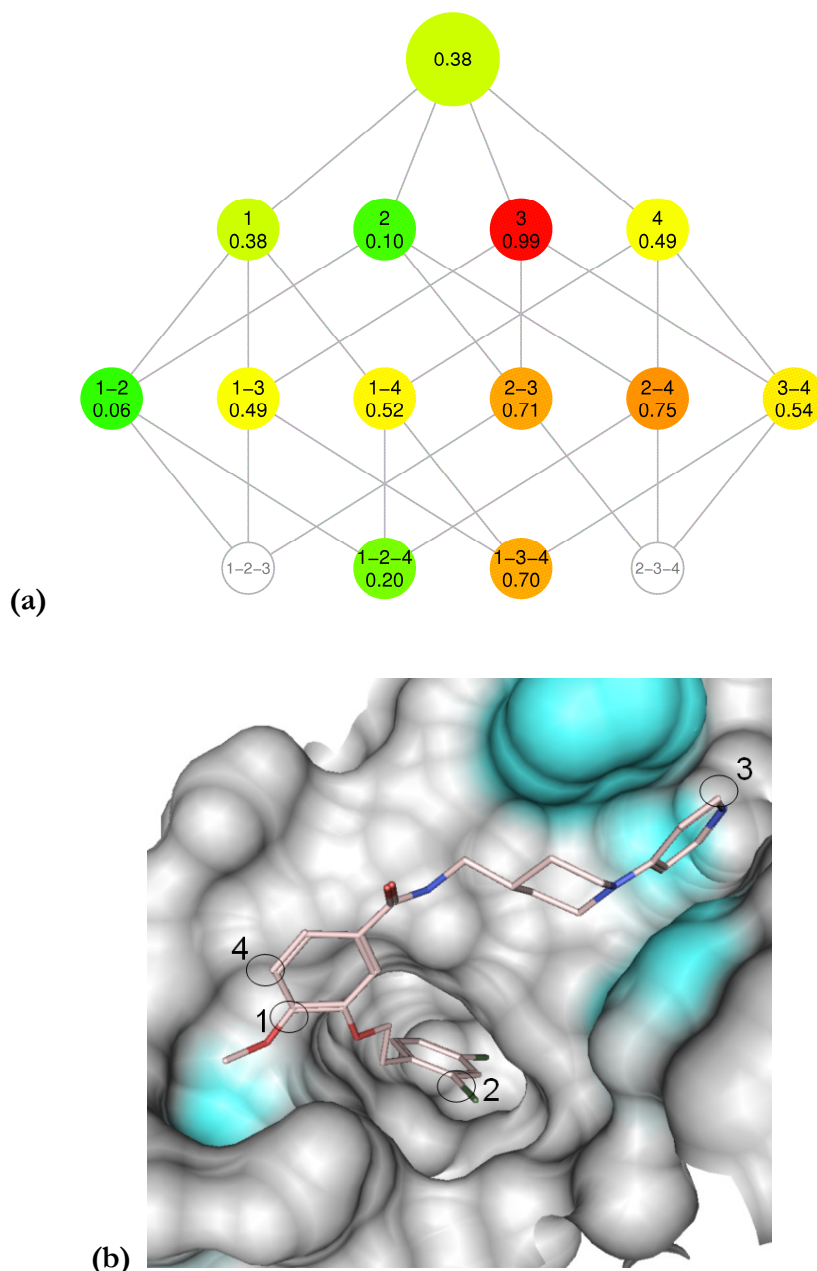


**Figure 3.3: Tie-2 kinase inhibitors.** (a) CAG for a set of eight inhibitors of Tie-2 kinase with substitutions at up to five different sites, (b) X-ray crystal structure of Tie-2 kinase in complex with inhibitor 14983 (PDB code 2P4I). In this structural representation, the active site of the enzyme is depicted as a solid surface. Selected hydrophobic pocket is highlighted with cyan coloring and residue Ala905 is labeled for reference. The inhibitor is shown in stick representation using the following atom color codes: light-rose for carbon; red for oxygen; blue for nitrogen and green for fluorine. Substitution sites in the inhibitor are circled and labeled accordingly.

However, in compounds corresponding to node 2-4, the methyl position is exchanged between sites 2 and 4, which resulted in moderate SAR discontinuity. This is the case because as can be seen in the complex (Figure 3.3b), the hydrophobic pocket facing site 4 is smaller than the one facing site 2 and the methyl group at site 4 is likely to form an unfavorably close contact with the Phe296 side chain (not shown), which reduced the compound potency by an order of magnitude.

Moreover, in compounds corresponding to node 2-3-4, the methyl position is also switched between sites 2 and 4, and site 3 contains a piperazine or morpholine group or no substituent. These modifications at site 3 are also present in compounds corresponding to node 2-3 but only lead to moderate SAR discontinuity. However, simultaneous modifications at sites 2, 3 and 4 act in concert and lead to a considerable degree of SAR discontinuity. Nodes 3-5 and 3-4-5 also display moderate and high discontinuity, respectively, which primarily results from the change of a cyclic to an acyclic substituent at site 3 (the detailed R-groups are given in Appendix C). Preferences for substituents at site 3 are not apparent from the structure, except that crystallographic temperature factors indicate significant protein backbone flexibility in the region, which might give rise to induced fit effects.

**Factor Xa, series 1:** This series contains 20 analogs that differ at four substitution sites and span a large potency range of 13 nM – 1053 nM. In this case, substitution site 3 forms a prominent SAR hotspot (Figure 3.4a). SAR discontinuity at this site is largely determined by the presence or absence of a hydroxyl substituent at the pyridine moiety that is engaged in strong  $\pi$ - $\pi$  stacking interactions with aromatic binding site residues (Figure 3.4b).



**Figure 3.4: Factor Xa inhibitor series 1.** (a) CAG for a set of 20 analog factor Xa inhibitors with four substitution sites, (b) Crystal structure of inhibitor 13664 bound to factor Xa (PDB code 2BMG). In this structural representation, the active site of the enzyme is depicted as a solid surface. Selected hydrophobic sub-sites are shown with cyan surface coloring. The inhibitor is shown in stick representation using the following atom color codes: light-rose for carbon; red for oxygen; blue for nitrogen and dark-green for chlorine. Substitution sites in the inhibitor are circled and labeled accordingly.

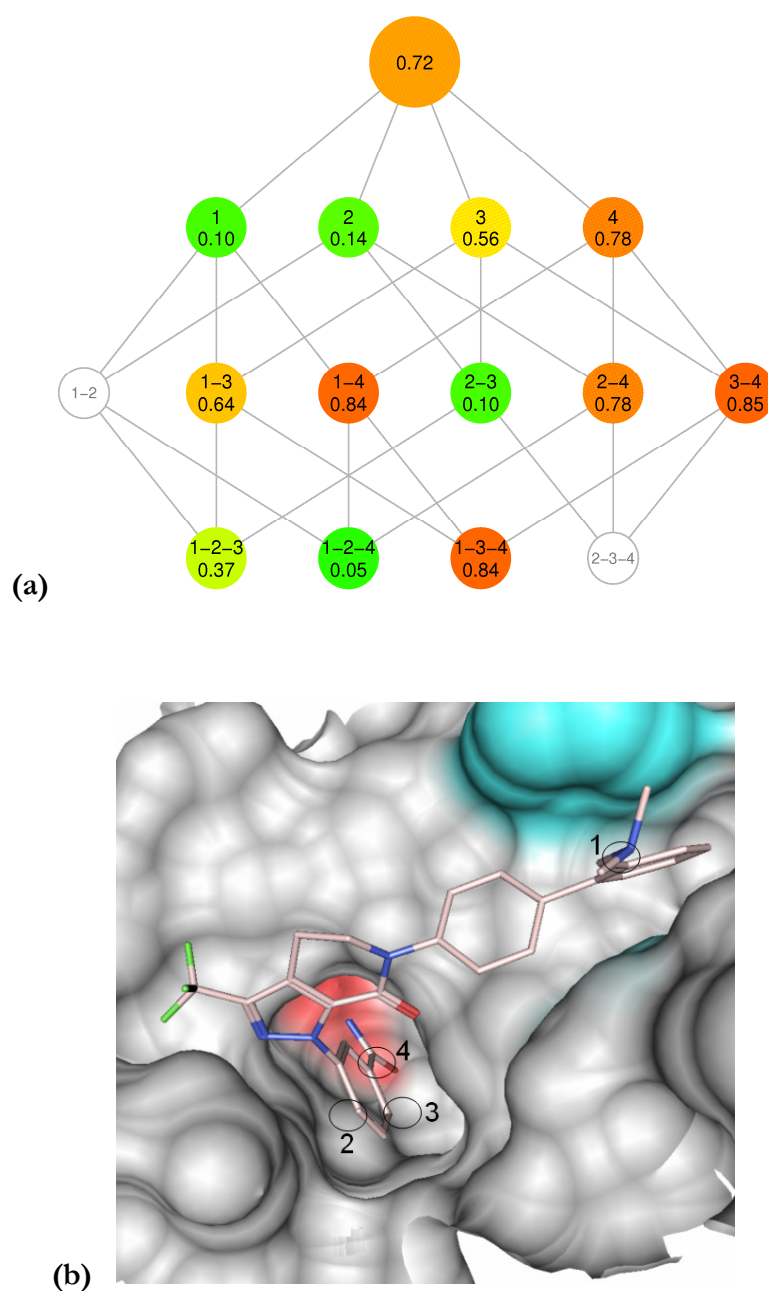
Addition of the OH-group at position 3 to the pyridine group enables tautomerization, which destabilizes its  $\pi$ -electron system and diminishes the stacking interaction. A weakly potent analog having a bromine substituent at site 4 also introduces moderate SAR discontinuity in this compound subset because a bromine at site 4 is expected to be larger at the entrance of the small



hydrophobic sub-pocket harboring site 1 substituents (Figure 3.4b) and imposes steric constraints. Hence, nodes including substitution sites 3 and 4 make the overall largest contributions to SAR discontinuity.

**Factor Xa, series 2:** This alternative inhibitor series consists of 13 highly potent analogs (potency values ranging from 0.18 nM to 88 nM) that differ at four substitution sites. The resulting CAG, presented in Figure 3.5a, mirrors SAR features that are significantly different from the factor Xa inhibitor series discussed above (factor Xa, series 1). Series 2 displays a well-defined pattern of compound subsets representing discontinuous SARs all of which involve substitutions at site 4 that result in potency differences of up to 2 orders of magnitude. Figure 3.5b shows the X-ray structure of the complex containing the inhibitor that has the most favorable substituent at this position.

The inhibitor-enzyme complex confirmed that site 4 substituents reach into the specificity determining S1 pocket in the active site of factor Xa that contains the residue Asp189 at the bottom. Interactions with this residue (or corresponding residues) represent a critical activity cliff and are a hallmark of trypsin-like serine protease inhibition.



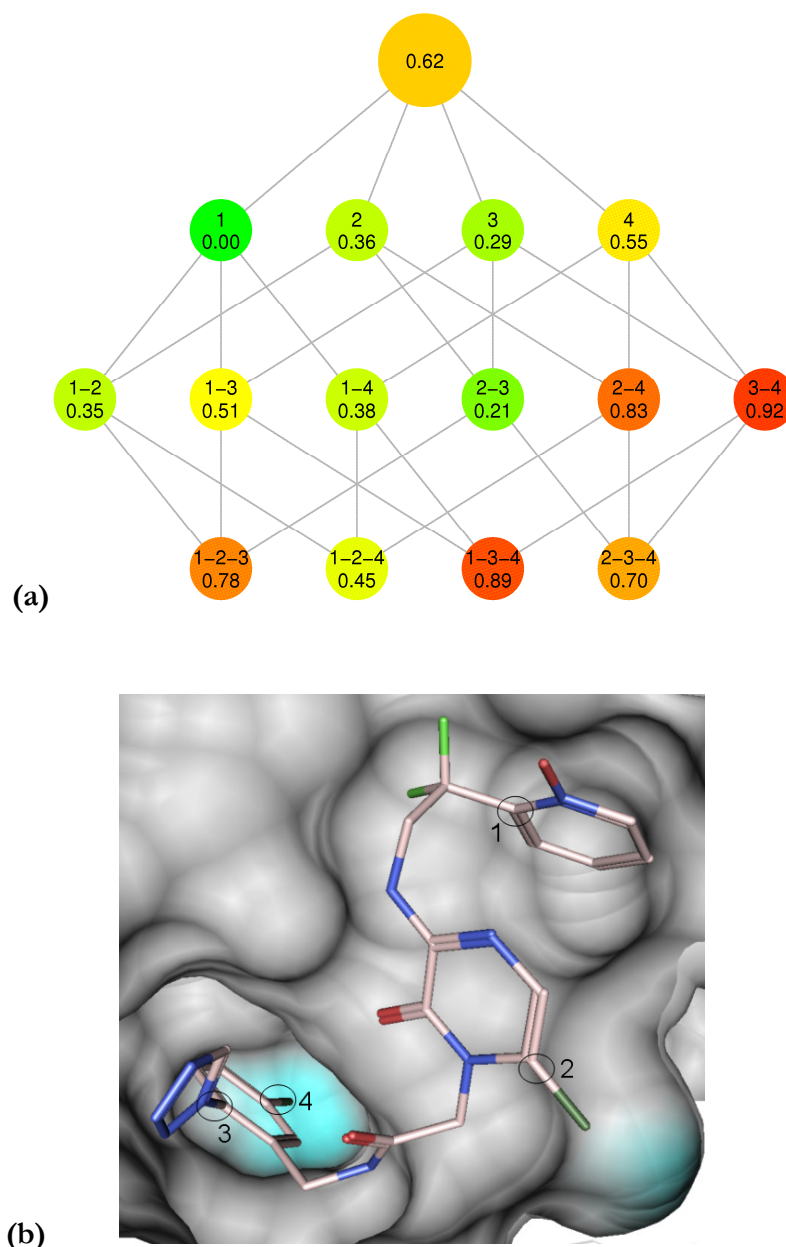
**Figure 3.5: Factor Xa inhibitor series 2.** (a) CAG for another analog series consisting of 13 factor Xa inhibitors with four substitution sites, (b) Complex crystal structure of factor Xa with inhibitor 12733 (PDB code 2G00). In this structural representation, the active site of the enzyme is depicted as a solid surface. Selected hydrophobic sub-site is shown with cyan surface coloring and charged/polar pockets with red coloring. The inhibitor is shown in stick representation using the following atom color codes: light-rose for carbon; red for oxygen; blue for nitrogen and light-green for fluorine. Substitution sites in the inhibitor are circled and labeled accordingly.

The inhibitor shown in Figure 3.5b fills this pocket and forms a hydrogen bonding network with residues Asp189 and Gly218, which is consistent with its high potency. Compared to this interaction constraint,

substitutions at other sites or site combinations introduce considerably less SAR discontinuity.

**Thrombin:** This set of thrombin inhibitors contains 13 analogs having four substitution sites and potency in the range of 0.0015 nM - 940 nM. In this case, most significant SAR discontinuity is observed for combinations of modifications involving sites 2, 3, and 4 (Figure 3.6a). Among individual sites, only substitutions at site 4 lead to notable discontinuity, which is due to an analog having a chlorine substituent at site 4 (in the absence of a chlorine at site 2). The simultaneous addition of chlorine substituents at both sites (node 2-4) further increases SAR discontinuity by causing a potency increase of several orders of magnitude. The strong discontinuity at node 3-4 is mainly due to variations at site 3 (triazole, tetrazole, or no substituent). Analogs at nodes 1-2-3, 1-3-4 and 2-3-4 combine the modifications described above that strongly affect potency.

The X-ray structure in Figure 3.6b contains the most potent analog having chlorine substituents at both sites 2 and 4. Substituents at site 3 are partly solvent exposed and the significant SAR discontinuity introduced by site 3 variations reflected in the CAG is difficult to explain in terms of interactions seen in the X-ray structure. Site 4 points into the S1 pocket and interactions involving key residues within the S1 pocket represent a pronounced activity cliff.



**Figure 3.6: Thrombin inhibitors.** (a) CAG for 13 analog thrombin inhibitors with four substitution sites, (b) Crystal structure of inhibitor 34 bound to the active site of thrombin (PDB code 1SL3). In this structural representation, the active site of the enzyme is depicted as a solid surface. Selected hydrophobic spots are shown with cyan surface coloring. The inhibitor is shown in stick representation using the following atom color code: light-rose for carbon; red for oxygen; blue for nitrogen; dark-green for chlorine and light-green for fluorine. Substitution sites in the inhibitor are circled and labeled accordingly.

In this analog series, however, the large specificity determining S1 pocket is not occupied with a positively charged group but with a chlorine atom site 4 substituent that strongly interacts with the  $\pi$ -electron system of Tyr228. The chlorine substituent at site 2 also fills the S2 hydrophobic pocket and the simultaneous presence of both chlorine substituents leads to a highly potent analog. This results in a strong SAR discontinuity which is clearly detectable in

the CAG (Figure 3.6a). The cooperative nature of these two sites is again hard to reconcile on the basis of the structure. Thus, taken together, the strong SAR discontinuity introduced by 2-4 and 3-4 substitutions in this series of thrombin inhibitors, as revealed by CAG analysis, could not have been deduced from interaction patterns in the X-ray structure.

The primary motivation of this study was to analyze SAR information contained in analog series and interpret the results at the level of protein-ligand interactions seen in complex X-ray crystal structures. The outcome helps to better understand how SAR discontinuity detected at the ligand level is reflected by interaction information derived from complex crystal structures and, if possible, arrive at a structural interpretation of individual activity cliffs. The systematic extraction of SAR information from compound series is a non-trivial task. For this purpose, we have developed combinatorial analog graphs that, on the basis of R-group decomposition and SARI scoring, make it possible to organize compound sets and identify substitution patterns that are responsible for activity cliffs and induce SAR discontinuity. These graph representations have been applied here to analyze selected analog series directed against different targets. Care has been taken to select series for which X-ray structural information was available and that shared large maximum common subgraphs and multiple substitution sites. This explains why the analog series studied here were of relatively small size because not very many analogs could be identified that met these requirements. However, the selected series were well-analyzed to study SAR discontinuity resulting from minor variations of substitutions in otherwise identical molecules. For each of these series, compound subsets with well-defined substitution patterns were identified that introduced significant degrees of SAR discontinuity and in a number of cases, activity cliffs could be readily mapped in X-ray structures. However, results also indicated that discontinuity patterns substantially differed between analog series, as exemplified by the two factor Xa inhibitor series, and that it was not possible in all cases to rationalize SAR determinants at the structural level. Thus, although often assumed, there is not always a consistent and close correspondence between SAR discontinuity and compromised receptor-ligand interactions. The relationship between SAR information encoded in analog series and receptor-ligand interactions seen in receptor-ligand structures is more complex. In some instances, individual SAR hotspots revealed in graph representations could be easily associated with critical interactions as, for example, in the case of the S1 site in thrombin or the tautomerization effects in factor Xa inhibitors. By contrast, the SAR features of, for example, substitution site combinations in thrombin inhibitors could not be rationalized in structural terms. On the other hand, structural analysis helped to clarify SAR ambiguities detected at the ligand

level as in the case of carbonic anhydrase II inhibitors. Thus, taken together, the results of this study also point at the complementary nature of LB SAR and structural analyses.

### 3.4 Summary

In study, SAR discontinuity information deduced from combinatorial analog graphs of different inhibitor series has been analyzed in light of receptor-ligand interactions in complex crystal structures, which has made it possible to explore activity cliffs at the small and macromolecular level. Although many effects of substitutions at defined sites in inhibitors could be rationalized in structural terms, SAR discontinuity detected in analog series could not only be attributed to the presence or absence of specific receptor-ligand interactions. However, structural interpretation helped to better understand the origin of SAR discontinuity in cases where LB analysis was insufficient. Clearly, information provided by systematic comparison of analogs and by analysis of complex crystal structures was highly complementary in a number of cases. Approaches for the extraction of SAR information from compound data sets should provide attractive starting points for the detection of activity cliffs and further exploration of local SAR patterns and SAR discontinuity at the target structural level.

# Chapter 4

## Identification of dual cathepsin K and S inhibitors

The previous part of the thesis, presented in chapters 2 and 3, was focused on analysis and utilization of protein-ligand interaction information for applications in VS and protein-ligand interaction-based SAR discontinuity analysis in analogue series. This part of the thesis, presented in chapters 4 and 5, focuses on practical applications of different VS methods for the identification of new inhibitors of selected cysteine proteases and a membrane-bound serine protease. Two major VS campaigns were carried out to identify dual cathepsin K and S inhibitors and matriptase-2 inhibitors. While cathepsins K and S are cysteine proteases, matriptase-2 is a newly identified type II membrane-bound serine protease. These proteases are considered to be important current pharmaceutical targets due to their involvement in bone resorption, immune response and iron metabolism, respectively. In the following two chapters, detailed VS methodologies and successful identification of inhibitors of the aforementioned enzymes are reported (Stumpfe et al., 2010; Sisay et al., *in revision*).

### 4.1 Introduction

In VS, computational methods are applied to search large databases for compounds having a desired biological activity using ligand (Bajorath, 2002)

and/or target structure (Shoichet, 2004; Kubinyi, 2007) information as input. In LBVS, computational methods are used to extrapolate from known active compounds and identify structurally diverse small molecules having similar biological activity, (Bajorath, 2002; Eckert and Bajorath, 2007) an objective often referred to as scaffold or lead hopping (Cramer et al., 2004). This study is focused in the application of a LB computational screening method to identify new inhibitors of two current pharmaceutical targets, cathepsins K and S.

#### **4.1.1 Cysteine proteases**

Cysteine proteases are implicated in a variety of human physiological processes and also form an essential component of the life cycle of a number of pathogenic protozoa and viruses. Cysteine cathepsins belonging to the papain-like subfamily represent the largest and best characterized group of cathepsins. They have attracted considerable interest over the past decade where many of the cysteine cathepsins are considered important drug targets (Brömme and Kaleta, 2002; Nägler and Ménard et al., 2003; Vasiljeva et al., 2007; Frizler et al., 2010). Among these enzymes the cathepsins K, S and L that are involved in bone remodeling, antigen presentation, and apoptosis, respectively, have attracted particular attention (Brömme and Kaleta, 2002; Yasuda et al., 2005; Vasiljeva et al., 2007). These proteases are highly similar homologues with mature enzyme sequence identities ranging from 56% to 60% (cathepsins K and L, 60%; S and K, 57%; and S and L, 56%) (Nägler and Ménard, 2003).

#### **4.1.2 Cathepsin K as a drug target**

Bone remodeling is a dynamic lifelong process where old bone is removed from the skeleton (resorption) and new bone is added (bone formation). Two groups of cells, osteoblasts which secrete new bone and osteoclasts which break bone down, are responsible for bone remodeling. As a result, bone is added where needed and removed where it is not required. Cathepsin K plays a critical role in the osteoclast-mediated degradation of collagen, the major component of bone matrix (Troen, 2004), and is predominantly expressed in osteoclasts that mediate bone resorption. It is capable of cleaving native type I collagen and other components of the bone matrix such as osteopontin and osteonectin (Blair and Athanasou, 2004; Stoch and Wagner, 2008). Accordingly, cathepsin K has become an attractive target for the development of drugs to treat osteoporosis and other disorders characterized by increased bone resorption (Blair and Athanasou, 2004; Stoch and Wagner, 2008; Zhao et al., 2009). Osteoporosis is a condition characterized by bone loss and microstructural deterioration that result in skeletal fragility and an increased risk in bone



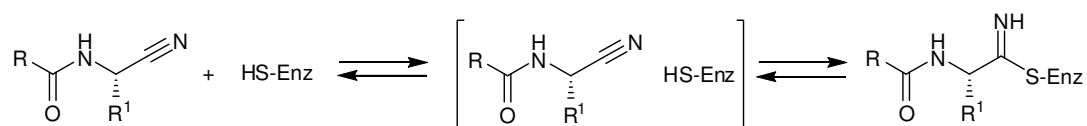
fractures. Moreover, cathepsin K is also implicated in rheumatoid arthritis and osteoarthritis (Brömme and Kaleta, 2002; Yasuda et al., 2005; Vasiljeva et al., 2007).

#### 4.1.3 Cathepsin S as a drug target

Cathepsin S is expressed by adipocytes and antigen presenting cells such as macrophages and B cells, and is involved in the control of antigen presentation by the major histocompatibility complex class II (MHC class II). MHC class II molecules present at the cell surface products of lysosomal proteolysis to T cells. MHC class II molecules are protected from inappropriate peptide loading during their maturation by association with the invariant chain. The invariant chain is progressively degraded until only a fragment, termed the class II-associated invariant-chain peptide (CLIP), remain to block the peptide-binding groove of MHC class II molecule. CLIP is later replaced by a diverse array of peptides derived from exogenous and endogenous proteins. Cathepsin S selectively degrades the MHC class II-associated invariant chain, which is a prerequisite for peptide antigen loading and presentation by the MHC class II complex (Riese et al., 1996; Driessen et al., 1999; Honey and Rudensky, 2003). Thus, cathepsin S has received much recent attention as a target for therapeutic intervention in a range of diseases of the immune system such as autoimmune and inflammatory disorders (Katunuma et al., 2003; Taleb et al., 2006; Leung-Toung et al., 2006).

#### 4.1.4 Cathepsin inhibitors

Cathepsin inhibitors are typically substrate analogues with an electrophilic "warhead". Most inhibitors discovered early on contained electrophilic warheads that react with the catalytic cysteine residue resulting in reversible or irreversible inhibition of the enzyme (Leung-Toung et al., 2006; Markt et al., 2008; Frizler et al., 2010). A nitrile represents a less-reactive functional group which is more desirable for therapeutic applications (Frizler et al., 2010). These nitrile-containing compounds interact with the active site cysteine residue forming a covalent reversible thioimide adduct (Figure 4.1). Recent investigations have revealed differences in electrophilic reactivity depending on the chemical environment of the cyano group (Oballa et al., 2007; MacFaul et al., 2009).



**Figure 4.1: Interaction of cysteine proteases with nitrile-based inhibitors.** Reversible formation of a thioimide adduct is illustrated.

Within the last years, a number of potent inhibitors of cathepsin K and S have been identified that contain less reactive electrophilic functionalities and inhibit via reversible, covalent interaction (Brömme and Kaleta, 2002; Yasuda et al., 2005; Leung-Toung et al., 2006; Löser et al., 2008; Löser et al., 2009; Frizler et al., 2010). Among the covalently interacting molecules, the first inhibitors of human cathepsin K with high selectivity, balicatib (Falgueyret et al., 2005) and odanacatib (Gauthier et al., 2008), have proceeded to clinical evaluations. Moreover, several inhibitors have also been reported that lack an electrophilic group and inhibit cathepsins non-covalently (Gustin et al., 2005; Leung-Toung et al., 2006; Tully et al., 2006c). Such non-covalent inhibitors having no reactive group are highly desirable as they would not react with unspecific nucleophiles leading to reduced side effects (Leung et al., 2000; Löser et al., 2010).

Computational screening methods have also been applied, but only in few cases, to identify cathepsin K or S inhibitors (Markt et al., 2008; Ravikumar et al., 2008; Stumpfe et al., 2009). This study is part of practical VS application efforts for the identification of cathepsin inhibitors aiming at (a) identifying inhibitors of both cathepsin K and S with previously unobserved scaffolds, possibly with non-covalent inhibition, which can be used as starting points for future chemical exploration and (b) further evaluating a new compound mapping algorithm, termed DynaMAD ('Dynamic Mapping to Activity class-specific Descriptor value ranges'). DynaMAD, is an unconventional VS method, for which, as of yet, only limited experience in practical applications is available. The characteristic feature of DynaMAD that sets it apart from other LBVS tools is that the method is designed to navigate molecular descriptor spaces of increasing dimensionality (whereas most compound classification techniques utilize low-dimensional reference space representations). Briefly, the DynaMAD is designed to map database compounds to activity-specific consensus positions in chemical space representations of step-wise increasing dimensionality (Eckert et al., 2006). The first step is the assignment of molecular descriptors to so-called dimension extension levels (DEL). The underlying idea is that descriptors that are most responsive to a biological activity represented by a reference set of known active compounds should only adopt very narrow value ranges. Descriptors are scored to account for the reference set specificity of their value ranges (scores range from 100 to 0). Then a score layer interval of 5 is applied so that a total of 20 DEL (0-20) are obtained. On the basis of their score,

different numbers of descriptors are assigned to each layer and hence the number of descriptors and the dimensionality of the descriptor reference space increase in a step-wise manner from layer to layer. In the next step, descriptor values are calculated for individual database compounds and they are mapped to the descriptor value ranges of the reference set at each layer. Only compounds whose values fall into the value range of each descriptor are retained for the next dimension extension step; the others are discarded. Hence, the number of database compounds decreases over the DEL until a final selection set remains.

## 4.2 Methodology

### 4.2.1 Data set and search strategy

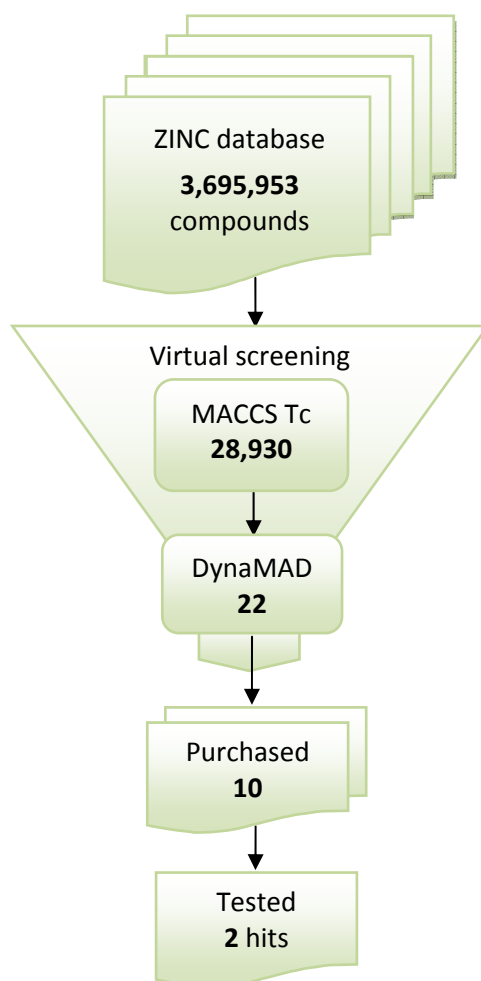
For the VS experiment, a reference set of 42 compounds from literature sources that inhibited both cathepsin S and K (usually with differential activity), covering a wide potency range from picomolar to micromolar values (potency range for cathepsin S: 200 pM - 0.2  $\mu$ M; for K: 140 nM - 100  $\mu$ M) were assembled. Nineteen of these 42 reference inhibitors contained the electrophilic nitrile group, whereas the others did not. The original literature sources of the reference set are provided in Appendix D.

As a source database,  $\sim 3.7$  million compounds of the publicly available ZINC database was used (Irwin and Shoichet, 2005). These compounds were first subjected to a molecular similarity-based pre-filtering step. A fingerprint consisting of the publicly available set of 166 MACCS structural keys (Durant et al., 2002) was used to search the database against each reference compound, and ZINC molecules were retained if they produced a Tc (Willett et al., 1998) value of not less than 0.75 compared to at least one of the reference molecules. For DynaMAD, a set of 155 2D molecular property descriptors (i.e. calculated from the molecular graph) was used which was readily available in the MOE suit of programs. This implies that the last mapping step was carried out in a 155-dimensional descriptor space.

## 4.3 Results and discussion

The MACCS similarity calculations yielded a ZINC pre-selection set of 28,930 compounds that were further subjected to DynaMAD analysis. Pre-filtering is not essential for DynaMAD calculations (that are computationally efficient), but the search was focused on database compounds that displayed at least some remote structural similarity to reference set molecules (considering that there is currently only limited structural information available about non-electrophilic cathepsin inhibitor chemotypes). During DynaMAD calculations, the 28,930

database molecules were reduced to only 22 candidate compounds that remained after the last of 20 dimension extension steps. Thus, the mapping calculations had high stringency and de-selected almost all database compounds. Out of the 22 candidate compounds, only 10 were available for purchase and could ultimately be acquired from commercial sources for further experimental evaluation. Figure 4.1 summarizes the VS and compound evaluation process. The structures of the 10 purchased candidate compounds and the ZINC IDs of the remaining commercially unavailable compounds is provided in Appendix D.

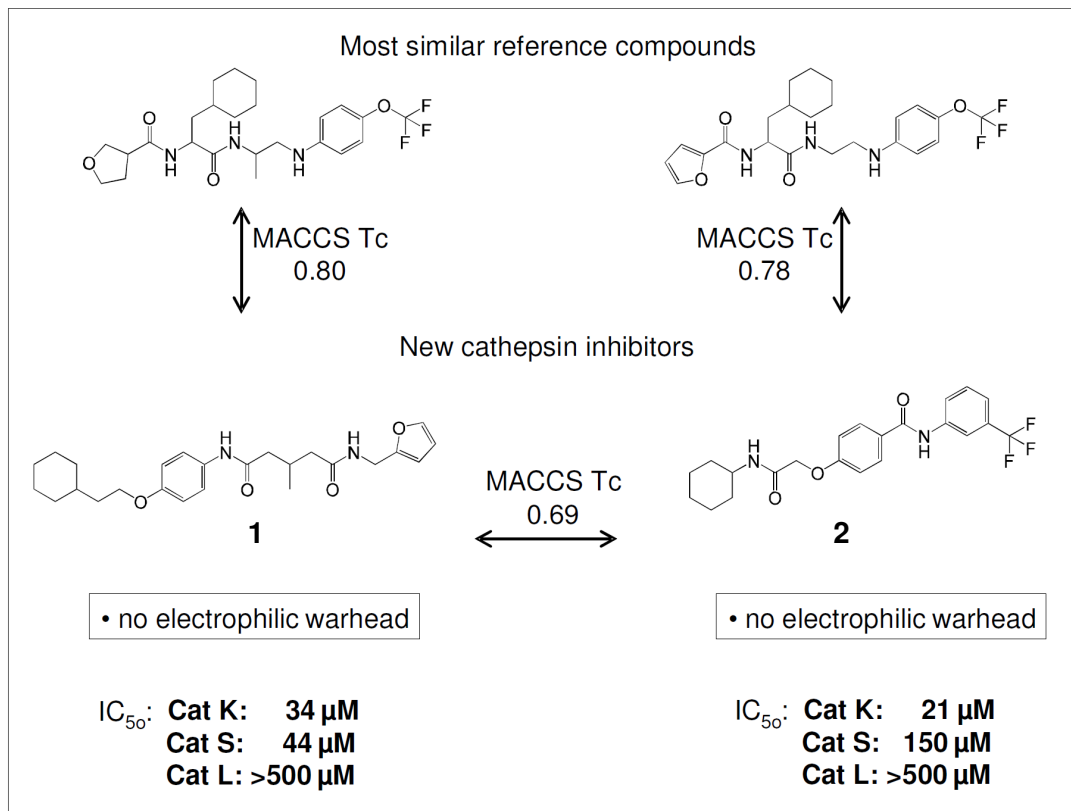


**Figure 4.1: Summary of the VS protocol.** The diagram summarizes the results of VS, compound selection, acquisition and testing.

These 10 compounds were tested for enzyme inhibition using a spectrophotometric assay for cathepsin S and L and a fluorometric assay for cathepsin K. Inhibition assays were performed by M. Frizler, Pharmaceutical Institute, University of Bonn. Assay details for each enzyme are provided in Appendix E.

The assay results showed that among the 10 candidate compounds, two were found to inhibit cathepsin K and S with  $IC_{50}$  values in the micromolar

range (which is often observed for structurally diverse hits identified by VS (Bajorath, 2002; Shoichet, 2004)). The structures of these compounds are shown in Figure 4.2 below together with their most similar reference compound.



**Figure 4.2: New cathepsin inhibitors.** Shown are the two newly identified compounds **1** and **2** that inhibited cathepsins K and S, but not L. On top, two of the most similar reference compounds are shown. The MACCS Tc values are also indicated for reference (the value between the two newly identified compounds was 0.69).

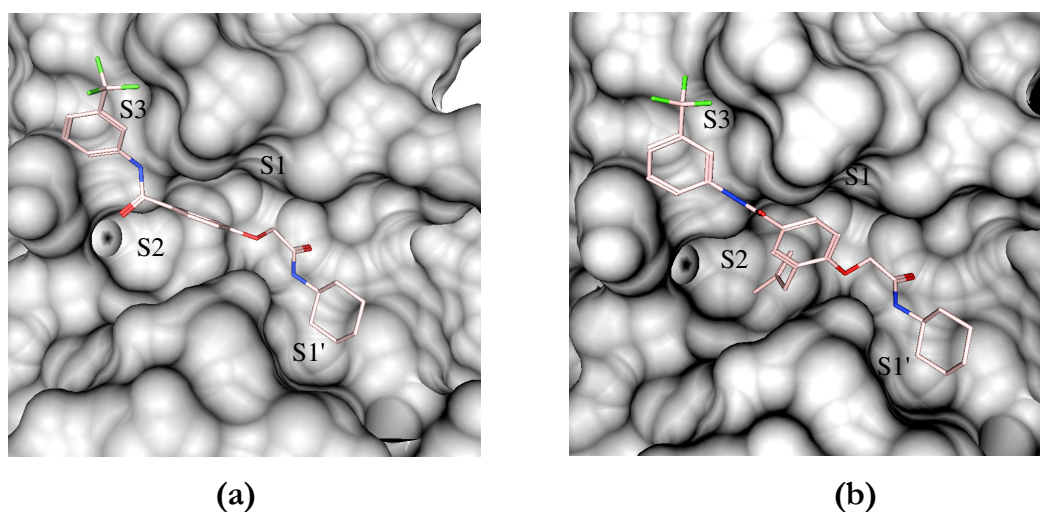
The compounds represent previously unobserved inhibitory scaffolds and are only remotely similar to the reference molecules. The MACCS Tc value to the most similar reference molecule is 0.80 for compound **1** and 0.78 for compound **2**. These values are lower than what would usually be expected for compounds having similar biological activity (Martin et al., 2002). Compound **1** has relatively comparable potency for cathepsin K and S (34 and 44 μM, respectively), whereas compound **2** inhibited cathepsin K (21 μM) about seven-fold more than S (150 μM), for which it is only a weak inhibitor.

Importantly, both inhibitors lack an electrophilic moiety such as a nitrile, although nearly half of the reference set contained such a group. Hence, these compounds add to the still limited spectrum of non-electrophilic cathepsin K and S inhibitors. These findings also demonstrate the ability of mapping calculations in high-dimensional descriptor spaces to abstract from eminent pharmacophore elements and detect new types of active compounds. Another

noteworthy feature of these two inhibitors is that they are not detectably active against cathepsin L, which is consistent with the applied VS strategy that did not take cathepsin L inhibitor information into account.

### 4.3.1 Binding mode analysis

To get an insight into the possible binding mode of compound **2** toward the active site of cathepsin K, molecular docking studies were performed using the FlexX (Rarey et al., 1996) docking software. The 3D structure of the compound was prepared in MOE. The active site of cathepsin K (taken from X-ray crystal structure with PDB ID 2ATO) was defined in FlexX taking residues within 6.5 Å around the crystallographic inhibitor. Docking was performed taking default values. The docking results (shown in Figure 4.3a) indicated that the compound most probably extends along the active site cleft with its trifluoromethylphenyl group located inside the S3 pocket forming  $\pi$ - $\pi$  interactions with Tyr67 and the cyclohexyl group inside the S1' site making several lipophilic contacts. In addition, the compound forms two hydrogen bonding interactions with the residues Asn161 and Gln19. The central phenyl ring is situated between the S1 and S2 pockets. Most known cathepsin K inhibitors bear a small aliphatic group, such as isobutyl side chain of leucine that occupies the hydrophobic S2 specificity determining pocket which is critical for tight binding into the enzyme active site (Altmann et al., 2003). In the case of compound **2**, it does not occupy the S2 pocket indicating a possible room for future optimization to improve its potency and selectivity. Therefore, from the model it can be hypothesized that modifying compound **2** with a group that extends deeper into the S2 site could improve potency and/or selectivity (Figure 4.3b). An attempt was done to modify the compound by attaching a substructure similar to the isobutyl group of leucine, i.e. a 2-methylallyl residue, on the central phenyl moiety. This residue would potentially extend into the S2 specificity pocket. The synthetic work was done by S. Dosa, Pharmaceutical Institute, University of Bonn. Unfortunately, a newly synthesized derivative did not show better inhibitory activity compared to the parent compound **2**. Further studies are required to clearly understand the binding mode and improve the potency of this compound.



**Figure 4.3: Binding mode of 2 in the active site of cathepsin K.** (a) Predicted binding mode of compound **2** in the active site of cathepsin K, (b) predicted binding mode of derivative of compound **2** with an additional moiety occupying the critical S2 specificity pocket. The enzyme is presented as surface and the inhibitor is shown in stick representation. The active site pockets are labeled.

## 4.4 Summary

In summary, by testing only 10 candidate compounds selected from a source database containing ~3.7 million molecules, two inhibitors of cathepsin K and S with new scaffolds have been identified in a VS application using the DynaMAD approach. These findings indicate that the mapping algorithm applied here detected inhibitory compounds in a more specific manner than often expected from LBVS methods. However, the identification of these cathepsin inhibitors is not only interesting from a methodological point of view because both inhibitors do not contain a nitrile or comparably reactive electrophilic groups. Therefore, they provide starting points for further chemical exploration of non-electrophilic prototypes of cathepsin K and S inhibitors. Molecular docking of compound **2** into the active site of cathepsin K indicated the presence of several putative intermolecular interactions.





# Chapter 5

## Identification of new matriptase-2 inhibitors

In the previous chapter, the results of a LBVS strategy for the identification of dual inhibitors of cathepsin K and S is presented. This chapter reports the results of a similar study but in this case focusing on the application of combined ligand and SBVS supported by knowledge-based compound design to identify inhibitors of matriptase-2. Matriptase-2 is a newly identified type II membrane-bound serine protease. It was very recently discovered that matriptase-2 plays a crucial role in body iron homeostasis by down-regulating hepcidin expression, which results in increased iron levels. Thus, matriptase-2 represents a novel target for the development of enzyme inhibitors potentially useful for the treatment of systemic iron overload (hemochromatosis). A comparative 3D model of the catalytic domain of matriptase-2 was generated in a previous project (Sisay et al., 2007) and was utilized for SBVS in combination with similarity searching and knowledge-based compound design (Sisay et al., *in revision*). Details of the search strategy used and results are discussed.

### 5.1 Introduction

#### 5.1.1 Type II membrane-bound serine proteases

The vast majority of known serine proteases are either secreted or sequestered in the cytoplasmic storage organelles awaiting signal-regulated release. Over the last years, a structurally distinct new class of serine proteases has been identified

that are transmembrane proteins containing an extracellular trypsin-like serine protease domain (Netzel-Arnett et al., 2003). These enzymes are involved in regulation of signal transduction between cells and their extracellular environment. They function in several important physiological processes such as digestion, cardiac function and blood pressure regulation, hearing, iron metabolism and epithelial homeostasis (Hooper et al., 2001; Szabo et al., 2003; Bugge et al., 2009; Choi et al., 2009). Even though the exact pathophysiological roles of many membrane anchored serine proteases remain to be elucidated, some numbers are indicated to be involved in different stages of cancer progression including growth, invasion, migration, and metastasis (Netzel-Arnett et al., 2003; Szabo et al., 2003; Noel et al., 2004; Lee et al., 2006; Szabo and Bugge, 2008; Bugge et al., 2009; Choi et al., 2009). The family of type II transmembrane serine proteases (TTSPs) possess a short intracellular N-terminal tail, a transmembrane domain and a large extracellular portion containing a variable stem region and a C-terminal serine protease catalytic domain of the chymotrypsin fold (Hooper et al., 2001; Szabo et al., 2003; Noel et al., 2004; Szabo and Bugge, 2008).

### **5.1.2 The matriptase subfamily**

The TTSPs can be divided into four subfamilies based on the phylogenetic analysis of the serine protease domain and the stem region. These include the hepsin/TMPRSS subfamily, the human airway trypsin-like protease (HAT)/differentially expressed in squamous cell carcinoma (DESC) subfamily, the matriptase subfamily and corin as the representative of a further subfamily. The members of the matriptase subfamily represent recently identified TTSPs with a unique stem composition and phylogenetically related serine protease domains. Known members of the matriptase subfamily are matriptase-1 (Shi et al., 1993; Lin et al., 1999), matriptase-2 (Velasco et al., 2002) matriptase-3 (Szabo et al., 2005) and the mosaic poly-protease, polyserase-1, (Cal et al., 2003) as well as its shorter splice-variant termed serase-1B (Okumura et al., 2006).

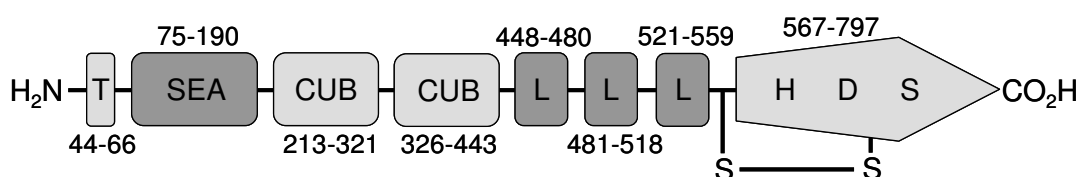
### **5.1.3 Matriptase-1 as a drug target**

Matriptase-1 (membrane-type serine protease 1, MT-SP1, suppressor of tumorigenicity 14), the most studied representative of the matriptase subfamily, was originally identified with a novel gelatinolytic activity in human breast cancer cells (Shi et al., 1993). By activating its potential substrates, e.g. pro-single-chain urokinase-type plasminogen activator (pro-uPA) and the proform of hepatocyte growth factor (HGF/scatter factor), matriptase-1 seems to play a relevant role in extracellular matrix degradation and cell scattering, whereas the

cleavage of profilaggrin is important for epithelial development (Lee et al., 2000; List et al., 2003; Lee et al., 2006; Szabo and Bugge, 2008). Matriptase-1 was shown to be overexpressed in a vast array of human tumors of epithelial origin including breast, prostate and ovarian cancers (Shi et al., 1993; Lin et al., 1999; Oberst et al., 2002; Lee et al., 2006; Uhlund, 2006) and has been implicated in tumor growth and metastasis in murine models of prostate cancer (Lin et al., 1999). Therefore, selective inhibition of matriptase-1 has therapeutic potential for the treatment of growth and metastasis of cancer.

#### 5.1.4 Matriptase-2 as a drug target

Matriptase-2, (TMPRSS6, transmembrane serine protease 6), was first identified in 2002 by Velasco and co-workers (Velasco et al., 2002) as a novel membrane-bound mosaic serine protease predominantly expressed in the liver. The extracellular stem region of matriptase-2 consists of a SEA (sea urchin sperm protein, enteropeptidase, agrin) domain-like region, two CUB (complement factor C1s/C1r, urchin embryonic growth factor, bone morphogenetic protein 1) domains and three repeats of low density lipoprotein receptor class A (LDLRA) domains (Figure 5.1) (Velasco et al., 2002; Netzel-Arnett et al., 2003; Szabo et al., 2003; Park et al., 2005; Ramsay et al., 2008).



**Figure 5.1: The modular structure of human matriptase-2.** Different domains are labeled as follows: T, transmembrane domain; SEA, sea urchin sperm protein, enteropeptidase, agrin domain; CUB, complement factor C1s/C1r, urchin embryonic growth factor, bone morphogenetic protein 1 domain; L, low density lipoprotein receptor class A domain. HDS indicates the catalytic triad (His, Asp and Ser) of the catalytic domain.

In contrast to matriptase-1, recent studies have shown that expression of matriptase-2 correlates with suppression of the invasiveness and migration of breast and prostate cancer cells (Parr et al., 2007; Sanders et al., 2008). However, precise functions of matriptase-2 in cancer remain to be further elucidated. Another interesting finding that recently attracted much attention is the correlation between mutations in the gene encoding matriptase-2 and iron-refractory iron-deficiency anemia (IRIDA), a condition that is poorly responsive to iron supplementary treatments (Du et al., 2008; Finberg et al., 2008; Melis et al., 2008; Ramsay et al., 2009). Iron is an essential trace element in mammalian metabolism and due to its generation of bio-reactive superoxide anions and hydroxyl radicals, levels of plasma iron require tight regulation (De Domenico

et al., 2008; Ramsay et al., 2009). Hepcidin, a small peptide hormone synthesized in the liver, is the homeostatic regulator of plasma iron levels and iron tissue distribution. It inhibits iron absorption from the intestine, regulates iron recycling and release from iron stores, and controls iron transport through the placenta. Hepcidin mediates the internalization and degradation of the iron exporter ferroportin, located on the surface of intestinal enterocytes, macrophages and hepatocytes, thereby inhibiting iron release into the plasma (De Domenico et al., 2008). As such, control of hepcidin expression represents a critical checkpoint for maintaining iron balance (Nemeth et al., 2004; Niederkofler et al., 2005). Matriptase-2 suppresses hepcidin expression (Du et al., 2008; Melis et al., 2008; Guillem et al., 2008; Folgueras et al., 2008) through proteolytic processing of cell surface hemojuvelin (Silvestri et al., 2008; Ramsay et al., 2009), a membrane-bound protein promoting hepcidin expression (Papanikolaou et al., 2004; Niederkofler et al., 2005; Malyszko, 2009). Due to the involvement in such a critical physiological process, matriptase-2 emerges as a new potentially important pharmaceutical target. Therefore, selective matriptase-2 inhibitors could be beneficial as pharmacological tools to further investigate its exact role in regulating iron homeostasis and might also be used for therapeutic intervention of frequent iron disorders such as systemic iron overload (hemochromatosis) or iron loading anemias where the level of hepcidin is inappropriately low.

Although peptide-based matriptase-2 inhibitors, such as aprotinin, have previously been reported (Velasco et al., 2002; Béliveau et al., 2009), small molecule matriptase-2 inhibitors have not been described so far. An essential advantage of small molecule inhibitors of matriptase-2 compared to large peptidic inhibitors, such as aprotinin, is their higher metabolic stability and easier synthetic accessibility for further developments. Therefore, this work was aimed at identification of new inhibitors of matriptase-2 by applying VS methods. A previously constructed high quality homology model of the catalytic domain of matriptase-2 (Sisay et al., 2007) was utilized for SB compound design, VS and subsequent evaluation. Initially, the model was used to interactively design four substrate-analog inhibitors, two amidino- and two chloro-substituted benzylamides. Following that, a VS campaign was carried out in the presence of these four compounds. On the basis of the screening calculations, the designed benzamidines were assigned a high priority with the ranked database compounds. These four substrate-analogue compounds were synthesized and their inhibitory profile was investigated in *in vitro* assays, which lead to the identification of the first low-molecular weight inhibitors of matriptase-2.

## 5.2 Methodology

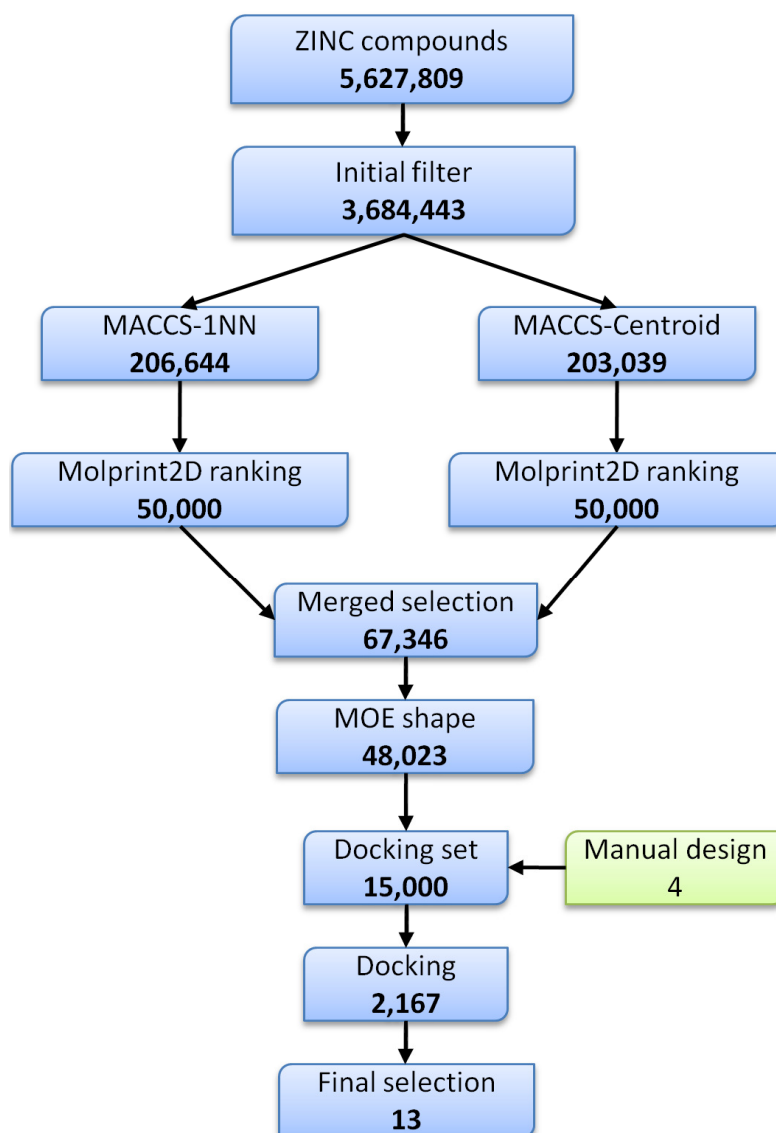
### 5.2.1 3D model of matriptase-2

Since an X-ray crystal structure of human matriptase-2 is currently not available, a detailed homology model of the catalytic domain of matriptase-2 was generated in a previous project (Sisay et al., 2007). The 3D structure was constructed using the crystal structure of the closely related matriptase-1 and other homologue enzymes using homology modeling, a procedure by which the coordinates of atoms of a target protein are predicted based on a topological sequence alignment of the target and a template protein(s) of known structure (Sali and Blundell, 1993).

### 5.2.2 Virtual Screening calculations

The VS protocol was carried out by applying both ligand and SB approaches sequentially. To reduce the number of database compounds to a reasonable number for molecular docking, initial property-based screening was applied followed by LB similarity searching by utilizing two different structural fingerprints methods, MACCS (Durant et al., 2002) and Molprint2D (Bender et al., 2004b). The MACCS fingerprint consists of a set of 166 annotated structural fragments which are used as keys to assess the level of similarity between a pair of compounds. On the other hand, Molprint2D is a circular atom environment fingerprint that generates varying numbers of strings depending on the complexity of the compound.

Following similarity searching, molecular docking was performed by applying two docking programs, DOCK6 (Meng et al., 1992) and FlexX (Rarey et al., 1996). The DOCK6 suite of docking programs is based on matching of spheres generated within the active site of the protein with ligand atoms and uses scoring grids to evaluate ligand orientations. FlexX uses a fast incremental construction algorithm which consists of base selection, placement complex construction and evaluation using a scoring function that estimates the free energy of binding. DOCK6 is more suitable for docking into large hydrophobic pockets whereas FlexX tends to be more efficient in docking of hydrophilic compounds in to active site pockets containing hydrogen bonding and charged groups. The detailed description of the VS and subsequent compound selection steps are given in Figure 5.2.



**Figure 5.2: Flow diagram showing ligand and structure-based virtual screening protocol.** The four manually designed compounds were additionally included into the 15,000 final docking set.

### 5.2.3 Ligand-based virtual screening

The ZINC public database (Irwin and Shoichet, 2005) containing a total of 5,627,809 compounds was initially filtered to remove compounds containing toxic and reactive groups by applying a broadened ‘rule of five’ (molecular weight: 200 - 600, logP: -2 - 6, donors: 1 - 10, acceptors: 1- 10, rotatable bonds: 0 - 18), reducing the number of compounds to 3,684,443. Then, eight known inhibitors of matriptase-1 (compounds 2, 8, 18, 20, 29, 31, 56 and 59 selected from Steinmetzer et al., 2006) were taken as a reference set for k nearest neighbor (1-NN) (Hert et al., 2004) and centroid (Schuffenhauer et al., 2004) similarity searching using MACCS structural keys (Durant et al., 2002) as a fingerprint and the Tc (Willett, 2005) as the similarity measure. Compounds

falling into the MACCS Tc interval between 0.6 and 0.8 were pre-selected in order to retain molecules with some structural resemblance to matriptase-1 inhibitors but omit analogs or other notably similar compounds; a total of 206,644 compounds from the nearest neighbor search and 203,039 from centroid search was obtained. Each of these pre-selections were re-ranked using the Molprint2D fingerprint search method (Bender et al., 2004), which is a higher-resolution similarity search tool than MACCS keys, and the top 50,000 compounds from each list were taken. Merging of the two ranking lists gave a total of 67,346 unique compounds. These database compounds were further ranked on the basis of an approximate shape matching procedure using the MOE relative to the known matriptase-1 inhibitors (with a cut-off value of 0.5) giving 40,023 compounds. The top 15,000 compounds were then considered for screening by molecular docking to which four knowledge-based manually designed compounds were added. The compounds were designed by Prof. Dr. T. Steinmetzer, Institute of Pharmaceutical Chemistry, University of Marburg.

#### 5.2.4 Structure-based virtual screening

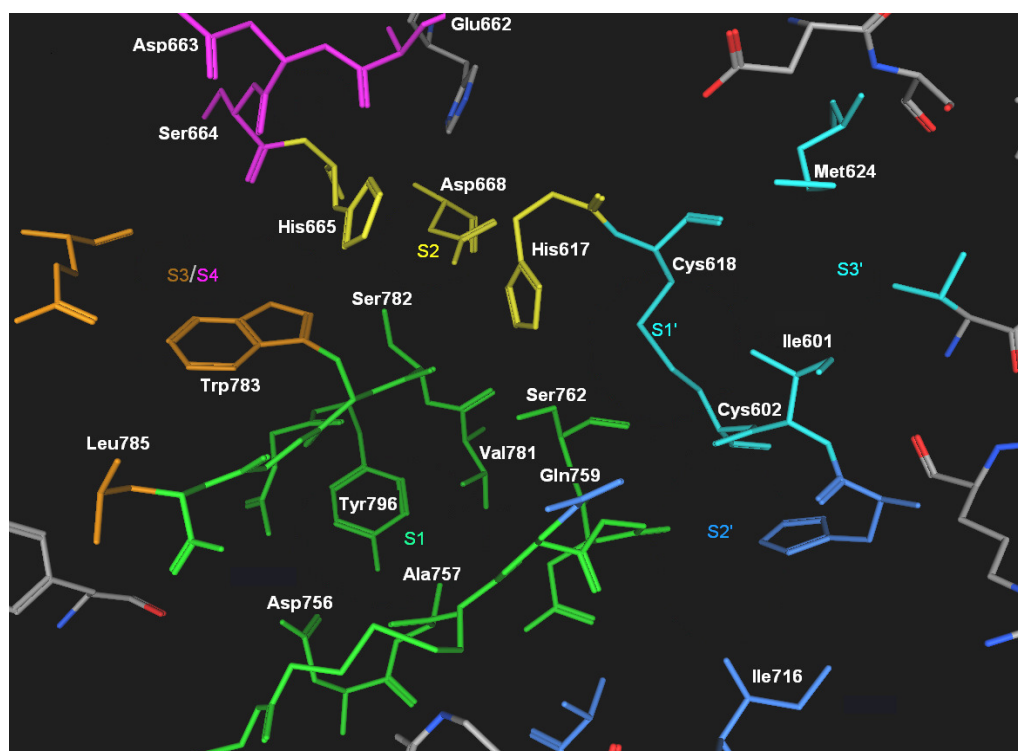
The homology model of matriptase-2 was used for docking. As docking programs, DOCK6 (Meng et al., 1992) and FlexX (Rarey et al., 1996) were applied for flexible ligand docking. For FlexX, the active site was prepared by taking all residues within 6.5 Å around a crystallographic inhibitor of matriptase-1 (PDB ID 2GV6) (Steinmetzer et al., 2006), after superposition of the matriptase-2 homology model and the X-ray crystal structure of matriptase-1. For DOCK6, a 10 Å radius was applied to generate and select spheres which are used to define and map the active site region. In both docking programs, docking parameters were adjusted by initially re-docking the inhibitor of matriptase-1 and reproducing its crystallographic pose to a reasonable level of accuracy.

After optimization of all the necessary parameters and preparation of ligands and active site of the enzyme, the 15,004 compounds were docked into the active site of matriptase-2 with an initial shape-based scoring followed by an energy-based evaluation scheme with the DOCK6 program. In the final DOCK6 ranking, a total of 2,167 compounds produced a DOCK6 energy score of less than -20 kcal/mol. These compounds were further considered for re-docking using the FlexX docking program. Finally, ten poses of the first 300 top-scoring database compounds were visually inspected and 13 compounds were selected as potential candidates for further testing and investigations. Two of the manually designed compounds were among the 13 compounds. The final compound selection was mainly based on (i) active site shape/chemical complementarity, (ii) occupation of the S1 specificity pocket by a basic group,

(iii) omission of compounds with solvent-exposed bulky hydrophobic groups, and (iv) detailed assessment of conformational strains.

## 5.3 Results and discussion

The active site of matriptase-2, shown in Figure 5.3, resembles that of trypsin-like serine proteases. At the bottom of the specificity determining S1 pocket, it has aspartic acid residue (Asp756) which is the major reason why matriptase-2 cleaves after a basic residue such as arginine. The S2 pocket is between His617(57) and His665(99) (residue numbering according to the original whole sequence of matriptase-2; matriptase-1 numbering is given in bracket for reference, Friedrich et al., 2002). The large S3/S4 pocket extends from Leu785(217) to the backbone carbonyl groups of Glu62(96), Asp663(97) and Ser664(98) (Sisay et al., 2007).

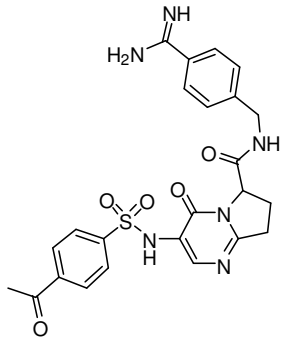
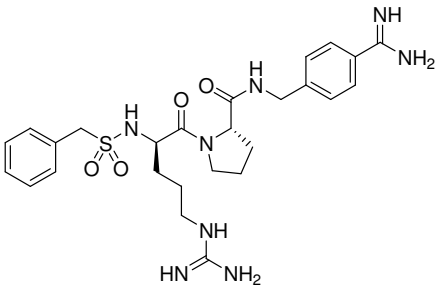
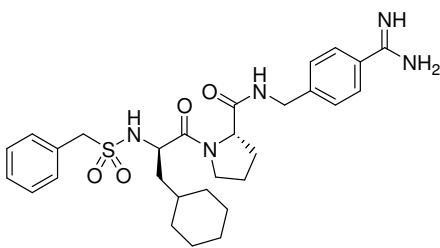
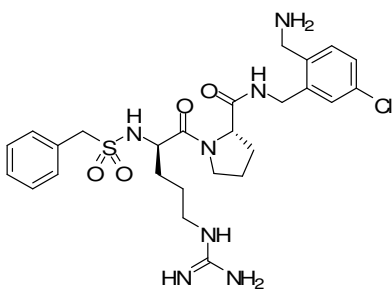
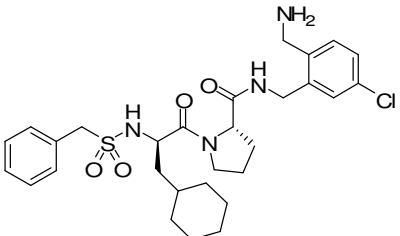


**Figure 5.3: Active site of matriptase-2.** Corresponding residues are labeled and color-coded according to the pocket they form. The numbering is based on the original whole matriptase-2 sequence. The model was taken from Sisay et al., 2007.

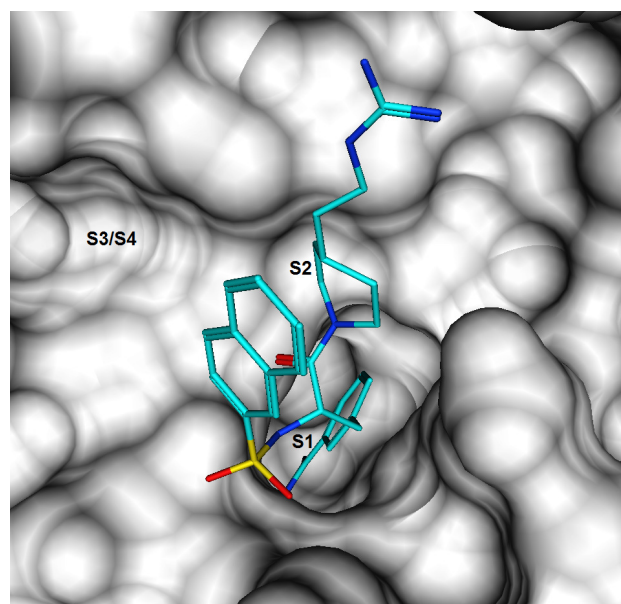
The structure of the four manually designed substrate analogue compounds and the best scoring database compound are shown in Table 5.1 below. The four compounds suggestions (compound **1-4**), originating from ‘knowledge-based’ design, which were included in the SBVS calculations were ranked at positions 2, 92, 6, and 402 respectively.



**Table: 5.1: Structure of selected compounds.** The four manually designed compounds, including the best scoring ZINC compound, are shown. 'Ranking' reports the rank of the compound in the final selection and the 'FlexX score' indicates the final docking energies in kJ/mol.

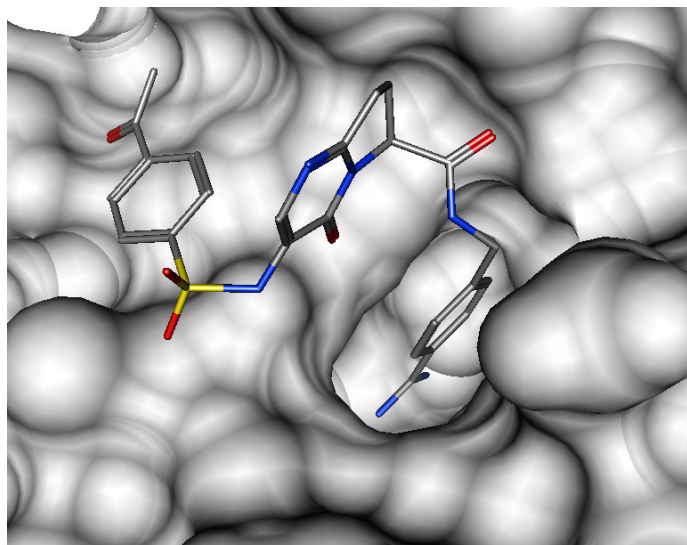
Ranking	Structure	Code	FlexX score (kJ/mol)
1		ZINC03838230	-46.59
2		1	-44.58
6		3	-41.21
92		2	-31.97
402		4	-25.40

The structures of known potent inhibitors of matriptase-1 (Steinmetzer et al., 2006), the X-ray structure of a matriptase-1/inhibitor complex (shown in Figure 5.4), and comparison of active site features, were taken into account to support the modeling efforts. The manual compound design was specifically based on the different features of the active site of matriptase-2. Accordingly, a basic benzamidine group was selected to potentially interact with the acidic Asp756(189) at the bottom of the S1 pocket. Proline was used as a linker which could occupy the small S2 pocket followed by a D-configured arginine expected to extend to the partially negatively charged portion of the upper part of the S3/S4 pocket. As an alternative, a D-configured cyclohexylalanine and chlorobenzylamine groups were selected instead of the benzamidine and D-arginine groups, respectively, resulting in four dipeptide amide-based compounds.



**Figure 5.4: The crystallographic matriptase-1 inhibitor complex (PDB ID 2GV6)** (Steinmetzer et al., 2006). The complex was used to deduce a likely orientation of putative matriptase-2 inhibitors in the binding site of the homology model of matriptase-2. The protein is presented as surface, the ligand as stick (with atom coloring scheme: cyan for carbon, red for oxygen, blue for nitrogen and yellow for sulfur) and the different active site pockets are labeled.

As an example, the docking pose of the best scoring database compound, with ZINC ID ZINC03838230, from the final selection set is shown in the Figure 5.5 below. From the docking pose it can be seen that the compound forms several favorable intermolecular interactions within the enzyme active site. It forms more than six hydrogen bonds and a salt bridge at the bottom of the S1 specificity pocket with Asp756(189) residue. The interaction surface properties of the active site of the enzyme in most part match the interaction properties of the compound. Therefore, this compound is expected to be a likely inhibitor of matriptase-2.



**Figure 5.5: Binding mode of the top scoring database compound (ZINC03838230).** Surface presentation of the active site pockets of matriptase-2 is shown. The ligand is presented as stick with atom type coloring scheme: gray for carbon, red for oxygen, blue for nitrogen and yellow for sulfur.

### 5.3.1 Enzyme inhibition assays

Unfortunately, none of the 13 prioritized compounds, including the manually designed ones, were available for purchase (the ZINC IDs of the database compounds are provided in Appendix D). To get experimental data on the prioritized compounds and evaluate the practicality of the screening strategy, the four manually designed compounds (dipeptide amides **1-4**, Table 5.1) were synthesized and tested against matriptase-2 and the homologous enzyme, matriptase-1. The synthetic work was done by Prof. Dr. T. Steinmetzer and M. Hammami, Institute of Pharmaceutical Chemistry, University of Marburg. As an enzyme source, the whole human matriptase-2 construct was cloned and expressed in human embryonic kidney (HEK) cells. Inhibition assays were performed using a shed form present in the supernatant of the cells that possesses the catalytic domain. Moreover, a purified form of the released enzyme was included in the study. Details of the enzyme expression and assay procedures are provided in Appendix E. The biological investigations were performed by Dr. M. Stirnberg, E. Maurer, S. Hauptmann, T. Bald and Stefan Frank, Pharmaceutical Institute, University of Bonn. The inhibitory potencies against matriptase-2 were compared with the activities against the catalytic domain of the structurally related matriptase-1 (Table 5.2). In addition, analysis of the binding mode of the compounds toward the two enzymes was performed.

**Table 5.2:** Summary of enzyme inhibition assays. Kinetic parameters for inhibition of human matriptase-2 and human matriptase-1 by the compounds 1-4 are given.

Compound.	$K_i \pm \text{SEM } (\mu\text{M})^a$		
	Human matriptase-2		Human matriptase-1
	in conditioned medium of HEK-MT2 cells	purified enzyme	recombinant enzyme
<b>1</b>	$0.19 \pm 0.01$	$0.17 \pm 0.02^b$	$0.055 \pm 0.003$
<b>2</b>	$> 10^c$	$> 10^c$	$0.22 \pm 0.01$
<b>3</b>	$0.29 \pm 0.02$	$0.46 \pm 0.06$	$0.77 \pm 0.15$
<b>4</b>	$> 10^c$	$> 10^c$	$2.1 \pm 0.3$

<sup>a</sup>  $\text{IC}_{50}$  values were determined from duplicate measurements with at least five different inhibitor concentrations.  $K_i$  values were calculated using the equation  $K_i = \text{IC}_{50} / (1 + [S]/K_m)$ .  $K_m$  values obtained for Boc-Gln-Ala-Arg-*para*-nitroanilide were  $210 \pm 7 \mu\text{M}$  for matriptase-2 in conditioned medium of HEK-MT2 cells,  $159 \pm 21 \mu\text{M}$  for purified matriptase-2, and  $381 \pm 33 \mu\text{M}$  for recombinant matriptase-1. SEM = standard error of measurement.

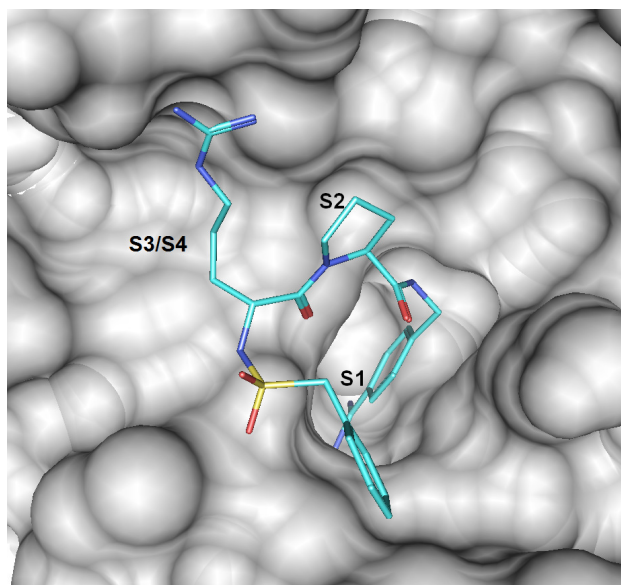
<sup>b</sup> triplicate measurement with five different inhibitor concentrations.

<sup>c</sup> duplicate measurement with three different inhibitor concentrations.

Compound **1** containing a D-arginine and a benzamidine moiety was the most potent inhibitor for both enzymes with a 3-fold higher potency for matriptase-1 ( $K_i = 55 \text{ nM}$ ) than for matriptase-2 in the conditioned medium ( $K_i = 190 \text{ nM}$ ) and purified matriptase-2 ( $K_i = 170 \text{ nM}$ ). Dipeptides with a 4-amidinobenzylamide group, such as **1**, are known potent inhibitors of thrombin and factor Xa (Schweinitz et al., 2006; Hellstern et al., 2007; Stürzebecher et al., 2007). The D-arginine and D-cyclohexylalanine derivatives **2** and **4**, both lacking the benzamidine moiety, had only marginal inhibitory activity against matriptase-2 ( $K_i > 10 \mu\text{M}$ ). Among the four dipeptide amides, only **3** exhibited a higher potency toward matriptase-2 in the conditioned medium ( $K_i = 290 \text{ nM}$ ) and the purified matriptase-2 ( $K_i = 460 \text{ nM}$ ) than against matriptase-1 ( $K_i = 770 \text{ nM}$ ).

### 5.3.2 Analysis of SAR and binding modes

The highly ranked compounds **1** (final docking rank 2) and **3** (rank 6) were found to be inhibitors of matriptase-2, but not compounds **2** (rank 92) and **4** (rank 402). Figure 5.6 shows the final manually optimized docking pose of the active compound **1**.

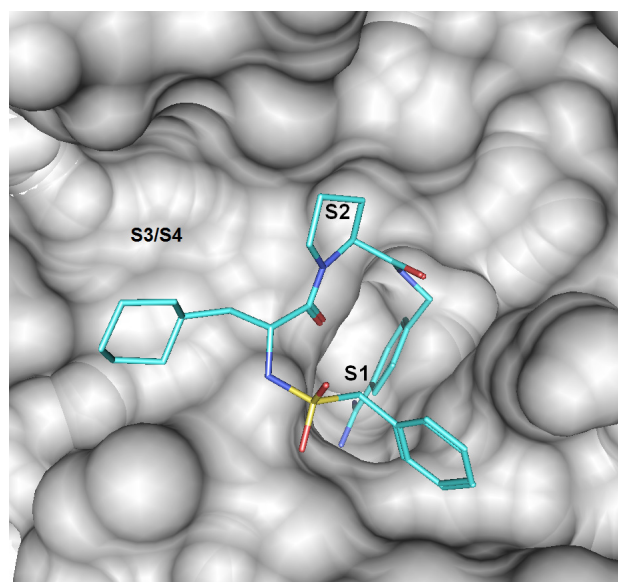


**Figure 5.6. Modeled enzyme-inhibitor complex.** Manually adjusted docking pose of matriptase-2/compound **1** complex is shown with surface presentation of the active site of the enzyme. The compound is shown in stick (coloring scheme: cyan for carbon, red for oxygen, blue for nitrogen and yellow for sulfur) and the different binding pockets are labeled for reference.

Due to accuracy limitations, it is generally not possible to predict detailed intermolecular interactions from docked poses, and here there is no attempt of doing so. However, likely structural characteristics of matriptase-2 inhibition can be explored on the basis of the model and binding mode of related inhibitors to related enzymes. As has been previously documented in several related enzymes, matriptase-2 has an acidic Asp group at the bottom of the S1 pocket. Accordingly, it can be seen from the optimized docking pose that the basic benzamidine moiety is well accommodated in the S1 specificity pocket with the amidine group forming a salt bridge with the acidic side chain of Asp756(189) at the bottom of the S1 pocket. Furthermore, in the docked pose, the guanidino group of arginine occupies the upper part of the S3/S4 pocket and is in hydrogen bonding distance to the carbonyl groups of Glu662(96), Asp663(97) and Ser664(98). In addition, the proline side chain binds to the S2 pocket. The inhibitor forms a short anti-parallel  $\beta$ -sheet to the backbone of Ser782(214) and Gly784(216) as has been experimentally shown in the binding of similar inhibitors to related enzymes (Schweinitz et al., 2004, 2006).

The position of the benzylsulfonamide moiety of the inhibitor could not be deduced with a high level of confidence. It resides either at the lower part of the S3/S4 pocket or just above the shallow hydrophobic subsite behind the S1 binding pocket, packing against the Cys758(191)–Cys787(220) disulfide bridge. The latter orientation would be supported by the X-ray crystal structure of the complex of factor Xa with a structurally related benzamidine inhibitor

where the benzylsulfonamide moiety was located in a corresponding subsite (Schweinitz et al., 2006). The binding of compound **3** likely resembles that of compound **1** (Figure 5.7) except that the cyclohexyl group would be positioned within the S3/S4 pocket for favorable interactions with Trp783(215) and Leu785(217).



**Figure 5.7. Modeled enzyme-inhibitor complex.** Shown is manually adjusted docking pose of matriptase-2/compound **3** complex. The enzyme active site is presented as surface and the different active site pockets are labeled. The ligand is shown in stick with coloring scheme: cyan for carbon, red for oxygen, blue for nitrogen and yellow for sulfur.

Interestingly, compound **1** was also a potent inhibitor of the closely related homolog enzyme, matriptase-1. When modeled into the binding site of matriptase-1, it became apparent that the accommodation of the benzylsulfonamide moiety would represent the major difference. In matriptase-1, this group is oriented toward the lower part of the S3/S4 pocket but could not extend into the shallow hydrophobic subsite behind the S1 pocket, because matriptase-1 has a tyrosine residue at position 146 whose side chain would block access of an inhibitor to this subsite. The guanidino moiety of **1** extends toward the upper part of the S3/S4 binding pocket forming hydrogen bonding to the Phe97 carbonyl oxygen and cation- $\pi$  interaction with Phe99 and Trp215, as has been similarly described for other inhibitors of matriptase-1 (Steinmetzer et al., 2009). The observed higher potency of compound **1** towards matriptase-1 is mainly because of the presence of the Phe97 phenyl ring system. This side chain together with Phe99 and Trp215 forms a solvent shielded  $\pi$ -electron system perfectly suited for a cation- $\pi$  interaction.

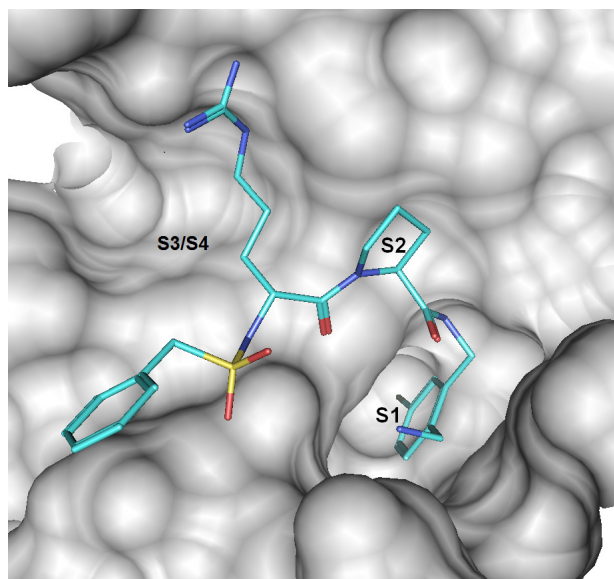
The replacement of D-arginine in **1** by D-cyclohexylalanine in **3** resulted in a stronger decrease in potency against matriptase-1 compared to matriptase-2. This finding might be rationalized by the reduced hydrophobic character of the

lower part of the S3/S4 pocket in matriptase-1 compared to matriptase-2. In matriptase-2, the cyclohexyl ring of **3** might favorably interact with Leu-785, but in matriptase-1, this interaction would be absent and the cyclohexyl ring extends towards the upper part of the S3/S4 pocket. In addition, in the case of matriptase-1, the loss of the cation- $\pi$  interaction contributes to the decreased activity (**1** *versus* **3**).

In compounds **2** and **4**, the benzamidine moiety is replaced by a *para*-chlorobenzylamine group. These compounds were inactive against matriptase-2. However, both compounds, in particular **2**, inhibited matriptase-1. Previously reported X-ray crystal structures of structurally similar inhibitors bound to thrombin (Rittle et al., 2003; Stauffer et al., 2005; Baum et al., 2009) showed that a chlorobenzene group occupied the S1 pocket with the chloro atom forming a van der Waals contact with Tyr228, which is conserved in matriptase-1 and -2. In a previous work focused on investigation of thrombin inhibitors containing a chlorobenzene group, Rittle *et al.* introduced an aminomethyl residue in *para*-position to the chlorine, as present in compounds **2** and **4**, and found further improvement on the inhibitory activity. In an X-ray structure of such a compound in the active site of thrombin, the amino group was observed forming salt bridge with Glu192 and a hydrogen bond to Gly216 at the entrance of the S1 pocket of the enzyme (Rittle et al., 2003). The major difference between matriptase-1 and matriptase-2 inside the S1 pocket is the presence of serine at position 190 in matriptase-1 instead of alanine in matriptase-2. Assuming a binding mode of compounds **2** and **4** in matriptase-1 similar to the one seen in thrombin (Rittle et al., 2003), which also has alanine at position 190, one would expect the benzamidine-chlorobenzylamine replacement to be tolerated by matriptase-2 rather than -1. However, compound **2** was identified as a selective inhibitor of matriptase-1, which can currently not be rationalized in structural terms. Additional inhibitors with structural variations at the P1 position will be required to better understand selectivity determinants between the two enzymes. From the X-ray structure of a related *N*-(3-chlorobenzyl)-prolinamide inhibitor (Baum et al., 2009) in complex with thrombin (PDB ID: 2ZC9) it can be inferred that compound **2** might bind similarly to matriptase-1 as the benzamidine-based inhibitor **1**.

However, while the amidine moiety of **1** would interact with Asp189 of matriptase-1, the chlorobenzylamine group of **2** would be buried inside the S1 pocket and the chloro atom would point towards the aromatic ring of Tyr228. Figure 5.8 shows a model representing this predicted putative binding mode.





**Figure 5.8: Modeled enzyme-inhibitor complex.** It shows a manually adjusted docking pose of matriptase-1/compound **2** complex. The enzyme active site is presented as surface, the ligand as stick (with coloring scheme: cyan for carbon, red for oxygen, blue for nitrogen and yellow for sulfur) and the different binding pockets are labeled.

## 5.4 Summary

In summary, in this study a combined SBVS and LBVS method, supported by knowledge-based design, was used to successfully identify the first low-molecular weight inhibitors of matriptase-2. The most potent compound **1** was, however, not selective for matriptase-2 over matriptase-1. Substitution of the guanidine moiety by cyclohexyl slightly reduced the potency against matriptase-2 (**1** *versus* **3**). This effect, however, is more pronounced in the case of matriptase-1 than matriptase-2, which might be due to the difference in hydrophobic character in the lower part of the S3/S4 pocket and the loss of the cation- $\pi$  interaction in matriptase-1. Replacement of the benzamidine by the chlorobenzylamine moiety (**1** and **3** *versus* **2** and **4**) resulted in a complete loss of inhibitory activity towards matriptase-2. Compound **2** was identified as an inhibitor with strong preference for matriptase-1 over matriptase-2. Such selectivity might be crucial in the development of matriptase-1 inhibitors as anticancer agents, because simultaneous inhibition of matriptase-2 is likely to have undesirable effects on body iron metabolism. However, the *N*-protected dipeptide amides **1** and **3** described herein can be used as leads to develop inhibitors with selectivity towards matriptase-2. For example, further exploring the P1 site in these compounds is expected to lead to a better understanding of selectivity determinants in the two matriptases. The selective inhibition of matriptase-2 may serve as a new potential strategy in the treatment of primary hemochromatosis and iron loading anemias.



# Chapter 6

## Inhibition and molecular modeling studies of brunsvicamides A - C against human leukocyte elastase

In the previous chapters, 4 and 5, different VS methods were applied to successfully identify dual cathepsin K and S, and matriptase-2 inhibitors. This chapter reports the identification of three analogous cyclic peptides, brunsvicamides A, B and C, as inhibitors of human leukocyte elastase (HLE) through enzyme inhibition assays. Subsequent molecular modeling studies were performed to get an insight into their possible binding mode (Sisay et al., 2009b).

### 6.1 Introduction

#### 6.1.1 HLE as a drug target

Human leukocyte (or neutrophil) elastase (HLE, EC 3.4.21.37) is a neutral protease which belongs to the chymotrypsin family of serine proteases. It has a primary specificity for small aliphatic residues, such as leucine, in the P1 position of the substrate. HLE is a major constituent in the azurophilic granules of human neutrophils and it is one of the many proteolytic enzymes released to combat invading foreign bodies during inflammation (Korkmaz et al., 2008). It is able to catalyze the cleavage of extracellular matrix proteins including fibrous elastin, an important extracellular matrix protein with unique property of elastic recoil and plays a major role in lung elasticity and proteolytic resistance. Under

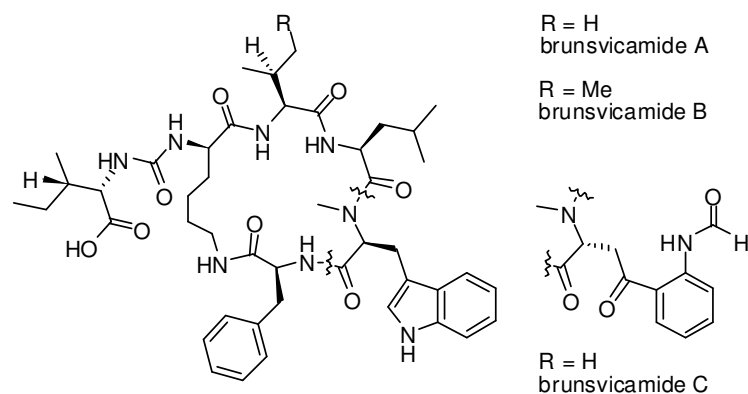
normal physiological conditions, the activity of HLE is regulated by endogenous inhibitors such as  $\alpha$ 1-protease inhibitor,  $\alpha$ 2-macroglobulin, and secretory leukocyte protease inhibitor but its excessive, uncontrolled activity could lead to tissue injury and several pathological states, including emphysema, chronic obstructive pulmonary disease, cystic fibrosis and rheumatoid arthritis (Chua and Laurent, 2006; Kodama et al., 2007). Due to its involvement in such pathophysiological processes, HLE has become an important pharmaceutical target particularly for the treatment of emphysema. Therefore, potent and selective HLE inhibitors can be used to reduce or treat HLE-mediated inflammatory disorders (Taggart et al., 2005; Pham, 2006; Siedle et al., 2007).

### 6.1.2 Cyclic cyanobacterial peptides

In recent years, several cyanobacterial secondary metabolites have been identified belonging to structurally novel cyclic peptides and depsipeptides (Moore, 1996; Sarabia et al., 2004). These natural products possess an attractive molecular architecture with a constrained conformation. They have an increased metabolic stability and display a variety of biological effects. Many of them inhibit enzymes, for example, peptides of the microcystin class (Honkanen et al., 1990; MacKintosh et al., 1990; Mehrotra et al., 1997) and oscillamides B and C (Sano et al., 2001) were found to inhibit protein phosphatases, while the cyanopeptolins, scyptolin A and B, (Matern et al., 2003) insulapeptolides A-D, (Mehner et al., 2008) and the anabaenopeptins B and F (Bubik et al., 2008) are inhibitors of elastases. Anabaenopeptins G, H, (Itou et al., 1999) I, J, and T (Murakami et al., 2000; Kodani et al., 1999) inhibit carboxypeptidase A. Trypsin and chymotrypsin inhibitory activities were reported for the cyanopeptolins A90720A and symplocamide A, respectively (Lee et al., 1994; Linington et al., 2008).

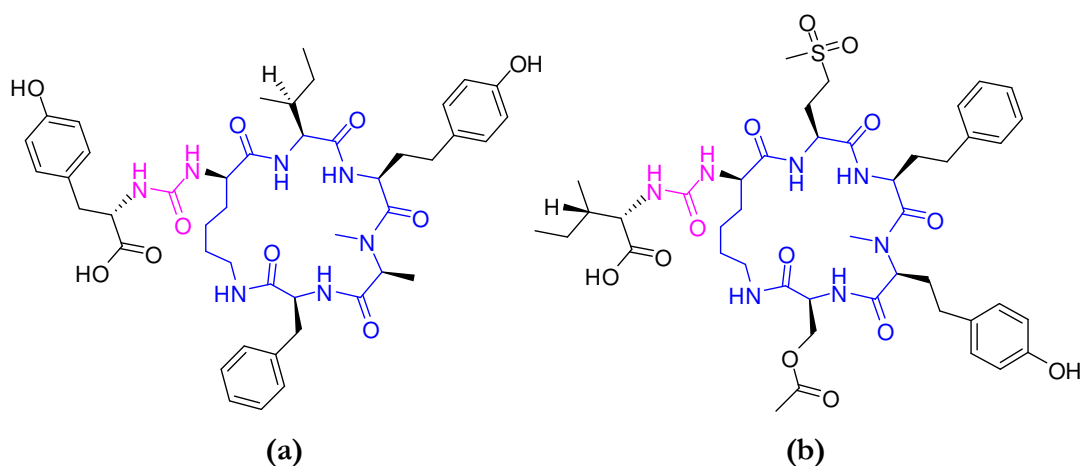
### 6.1.3 The brunsvicamides

The brunsvicamides A, B and C were originally isolated from the cyanobacterium *Tychonema* sp. (Müller et al., 2006; Walther et al., 2008) and are structurally related to the sponge-derived mozamides (Schmidt et al., 1997). They are characterized by having six amino acids, five of which form a 19-membered ring structure, closed by an amide bond between the carboxylic group of the C-terminal Phe and the  $\epsilon$ -amino group of the N-terminal D-Lys. The sixth amino acid is attached to the  $\alpha$ -amino group of the D-Lys via a urea moiety (Figure 6.1).



**Figure 6.1: Structures of the brunsvicamides A-C.**

The brunsvicamides, anabaenopeptins, oscillamides, (Marsh et al., 1997; Sano et al., 2001) and the nodulapeptins (Fujii et al., 1997) are structurally related by having a D-Lys-urea motif attached to a terminal amino acid and an *N*-methylated peptide bond in common, but differ in their amino acid sequence (Figure 6.2).



**Figure 6.2: Structures of the brunsvicamide related (a) anabaenopeptin A and (b) nodulapeptin A.** The two cyclic peptides are structurally similar to the brunsvicamides by having a D-Lys-urea motif attached to a terminal amino acid and an *N*-methylated peptide bond in common.

## 6.2 Methodology

The remarkable inhibitory properties of brunsvicamides against the tyrosine phosphatase B of *Mycobacterium tuberculosis* have been previously described (Müller et al., 2006), however, their protease inhibiting potential was not reported before. Based on previous findings that several cyclic cyanopeptides inhibit elastases, the brunsvicamides A-C were evaluated as potential inhibitors of HLE. Additionally, these cyanopeptides were assessed against a panel of proteases and two serine esterases including cathepsin G (which is also from human leukocytes), the serine proteases chymotrypsin and trypsin, as well as the cysteine protease cathepsin S. The two esterases, acetylcholinesterase (AChE)

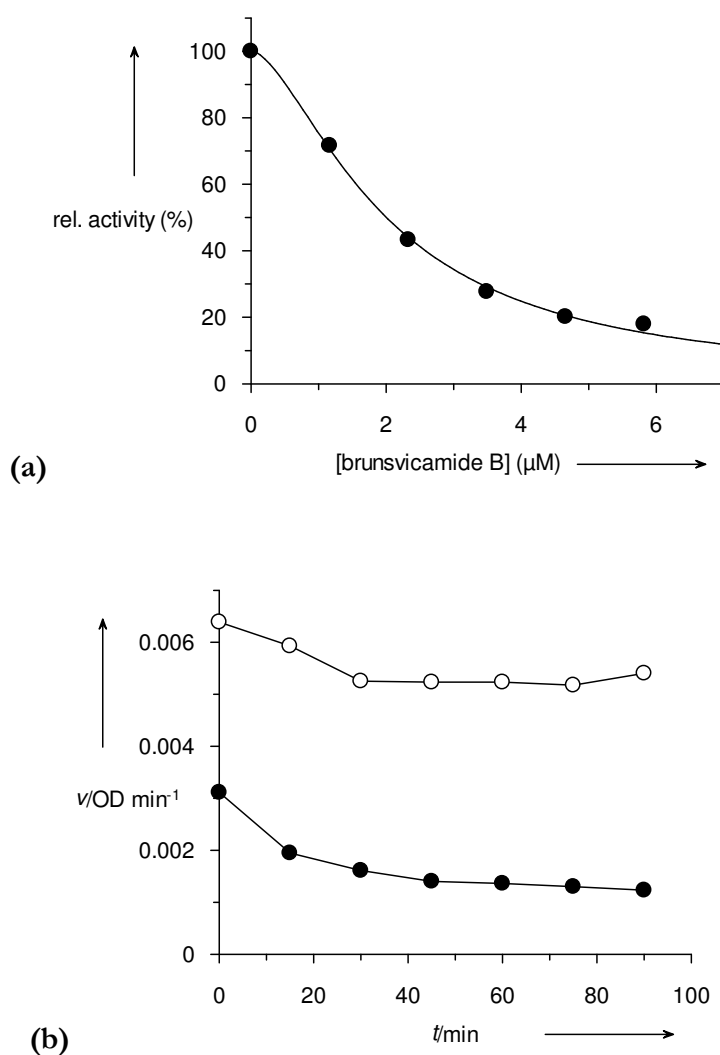
and cholesterol esterase (CEase), that share the acyl transfer mechanism with serine proteases, were also included in the study. Furthermore, molecular modeling studies were performed to get insight into the possible binding mode of the brunsvicamides in the active site of HLE.

## 6.3 Results and discussion

### 6.3.1 Enzyme inhibition assays

The three cyclic cyanobacterial peptides were tested for enzyme inhibition using a spectrophotometric assay system. The three compounds were kindly provided by Prof. Dr. G. König and Dr. C. Mehner, Institute for Pharmaceutical Biology, University of Bonn. The enzyme assays were carried out by S. Hauptmann, Pharmaceutical Institute, University of Bonn. For detailed description of assay and incubation experimental procedures see Appendix E. The concentration-dependent inhibition of HLE activity by brunsvicamide B is presented in Figure 6.3a. The progress curves of the HLE-catalyzed substrate consumption were linear over 10 min time course indicating time-independent inhibition by the brunsvicamides. The inhibition assay results with the corresponding  $IC_{50}$  values are given in Table 6.1. From the assay results, it can be seen that the brunsvicamides inhibited only HLE and not the related serine proteases or serine esterases. The brunsvicamides A-C were thus highly selective for HLE with  $K_i$  values of 1.1, 0.70, and 1.6  $\mu M$ , respectively, calculated assuming competitive inhibition. The potencies against HLE were comparable for all the three cyclic cyanopeptides.

Since the brunsvicamides are peptidic compounds, there is a possibility of degradation by HLE. To investigate the presence of any degradation of the brunsvicamides by HLE, incubation experiments were performed. HLE was incubated over 90 minutes with brunsvicamide C, and the enzyme activity was followed by adding aliquots to a chromogenic substrate (see Appendix E). A similar loss of HLE activity was observed in the presence and absence of the inhibitor during this incubation time period (Figure 6.3b). This indicates that brunsvicamide C was able to resist enzymatic hydrolysis by HLE and therefore, was not degraded.



**Figure 6.3: Inhibition and incubation experiments.** (a) Inhibition assay of brunsvicamide B against HLE. The assay was performed in the presence of  $100 \mu\text{M}$  ( $= 1.85 K_m$ ) of the chromogenic substrate MeO-Suc-Ala-Ala-Pro-Val-pNA. The data are mean values of duplicate measurements. The reactions were followed over 10 min, and the rates,  $v$ , were determined by linear regression. The rates in absence of inhibitor,  $v_0$ , were set to 100%. Nonlinear regression according to the equation  $v = v_0/(1 + ([I]/IC_{50})^x)$  gave a value  $IC_{50} = 2.00 \pm 0.08 \mu\text{M}$ . (b) Results of incubation experiments. Incubation experiments were performed in order to determine the activity of HLE in the presence (●) and absence (o) of brunsvicamide C. Final concentration of the substrate was  $100 \mu\text{M}$ , of brunsvicamide C was  $5.70 \mu\text{M}$ . The data are mean values of duplicate measurements.

**Table 6.1: Enzyme inhibitory activities of brunsvicamides A-C.** The calculated  $K_i$  values for HLE are given in brackets.

Enzyme	IC <sub>50</sub> [μM] <sup>a</sup>		
	Brunsvicamide A	Brunsvicamide B	Brunsvicamide C
HLE	3.12 ± 0.15 <sup>b</sup> ( $K_i$ = 1.1 μM)	2.00 ± 0.08 <sup>c</sup> ( $K_i$ = 0.70 μM)	4.42 ± 0.32 <sup>d</sup> ( $K_i$ = 1.6 μM)
Cathepsin G	> 100	100	81
Chymotrypsin	> 100	> 100	> 100
Trypsin	> 100	> 100	> 100
Cathepsin S	nd <sup>e</sup>	> 100	> 100
AChE	> 100	> 100	> 100
CEase	> 100	> 100	> 100

<sup>a</sup> Values with standard error were calculated from duplicate experiments at five different inhibitor concentrations, those without standard error are values or limits calculated from duplicate inhibition experiments at a single inhibitor concentration of 5 μg/mL (cathepsin S) or 25 μg/mL (other enzymes).

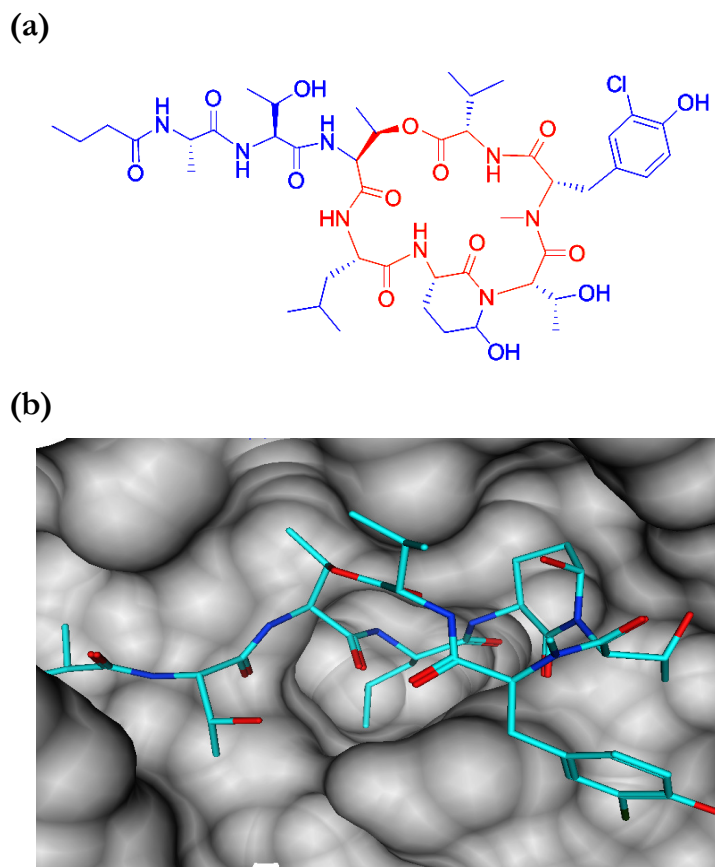
<sup>b</sup>  $x$  = 2.0. <sup>c</sup>  $x$  = 1.6. <sup>d</sup>  $x$  = 1.3.

<sup>e</sup> nd = not determined.

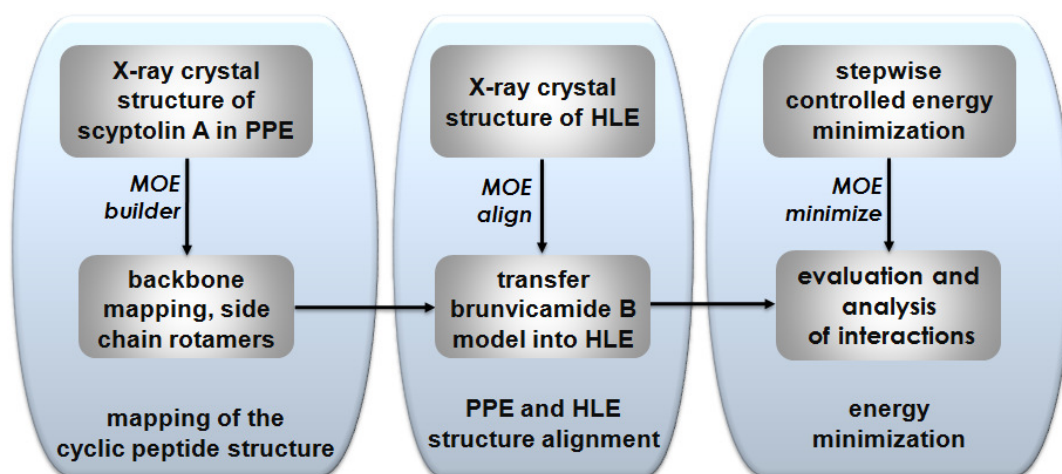
### 6.3.2 Binding mode analysis

To investigate the possible binding mode of the cyclic peptide inhibitors, molecular modeling was performed with the most potent representative, brunsvicamide B, in the active site of HLE. The initial approach was based on the X-ray crystal structure of scryptolin A in complex with porcine pancreatic elastase (PPE) (Matern et al., 2003). The cyclic peptide core of scryptolin A, formed by six amino acids, consists of 19 backbone atoms, which is exactly the same number as in the cyclic pentapeptide core of the brunsvicamides (Figure 6.4). Hence, we have investigated the possibility that the cyclic peptide core of scryptolin A might be mimicked by the brunsvicamides.

Computational modeling was carried out using the MOE software (Figure 6.5). First, the backbone structure of brunsvicamide B was mapped onto the 19-atom core of scryptolin A in its crystallographic conformation. After completing the backbone model, the corresponding amino acid side chains were replaced by close rotamer conformations, and residues outside the cyclic core were also added utilizing crystallographic backbone coordinates, to the extent possible.



**Figure 6.4: Structure of scyptolin A and PPE/scyptolin A complex.** (a) 2D structure of scyptolin A and (b) scyptolin A inside the active site of PPE is shown (taken from X-ray crystal structure PDB ID 1OKX, Matern et al., 2003).

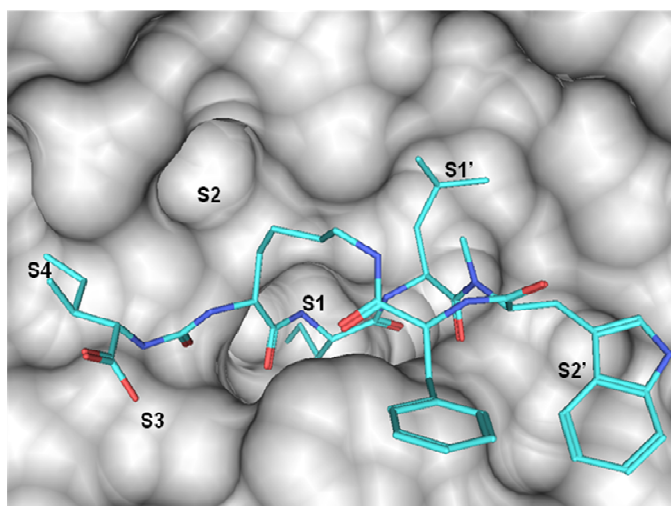


**Figure 6.5: Molecular modelling work flow.** The diagram summarizes the different steps involved in the modeling of brunsvicamide B inside the active site of HLE.

The modeled inhibitor mimicking the bound conformation of scyptolin A was then transferred into the binding site of the HLE crystal structure (Bode et al., 1986) after superposition of porcine pancreatic elastase and HLE X-ray

crystal structures. Intra- and intermolecular contacts of the resulting elastase-inhibitor model were optimized by a stepwise controlled energy minimization applying the Amber99 force field (Cornell et al., 1995) which is implemented in MOE.

Analysis of the model shown in Figure 6.6 revealed the presence of a number of plausible putative interactions. In the binding conformation analogous to scyptolin A, the two Ile residues of the inhibitor occupy the S3/S4 and S1 pockets, respectively, while the Leu residue occupies the S1' and the *N*-methyl-Trp occupies the S2' pocket. Only the S2 pocket remains unoccupied indicating a future optimization potential to improve the potency of the brunsvicamides A-C. Structural violations of the putative binding mode were not observed. These findings suggest that brunsvicamides A-C are capable of mimicking the bound conformation of scyptolin A and might hence act through a similar inhibitory mechanism.



**Figure 6.6: The model of brunsvicamide B bound to the active site of HLE.** The active site of the enzyme is shown as surface presentation and the ligand is shown stick.

Serine protease-catalyzed hydrolysis follows an acylation-deacylation mechanism. The brunsvicamides may act as alternate HLE substrates with a strongly decelerated deacylation step, thus leading to enzyme inhibition (Pietsch and Gütschow, 2002), however, the observed time independent inhibition of HLE by brunsvicamide B and the inability of HLE to regain its activity in the incubation experiment with brunsvicamide C, strongly suggest that the peptide inhibitor is tightly bound to the enzyme without getting degraded. These findings are in agreement with results on the inhibition of HLE by insulapeptolides (Mehner et al., 2008) and porcine pancreatic elastase by scyptolin A (Matern et al., 2003).

From the X-ray crystal structures of intact peptidic inhibitors bound to proteases, it has been thought that the rigidity of the enzyme-inhibitor



complexes effectively blocks the catalytic mechanism before the formation of the tetrahedral intermediate or before the formation of the acyl-enzyme. An alternative model suggests that the acyl-enzyme intermediate is formed readily, but that the peptide bond is more rapidly reformed, so that the intact form seen in crystal structures predominates (Zakharova et al., 2009).

In a recent study, Zakharova and coworkers (Zakharova et al., 2009) showed that a peptidic trypsin inhibitor binds to the enzyme like a normal substrate but resists hydrolysis by forming an equilibrium between cleavage and reformation of the scissile bond. Using X-ray crystallography, they studied the interaction between cleaved bovine pancreatic trypsin inhibitor (BPTI) with active and inactive forms of rat trypsin. A high-resolution (1.46 Å) crystal structure of a complex formed between a cleaved form of BPTI with a catalytically inactive rat trypsin variant showed that the inhibitor remains cleaved and the N- and C-terminal moieties of the cleaved bond were ideally positioned in the active site for resynthesis. This structure defines the positions of the newly generated amino and carboxyl groups following the acylation and deacylation step in the hydrolytic reaction. On the other hand, incubation of the cleaved BPTI with active rat trypsin resulted in the reformation of the scissile bond.

Close analysis of the structural complex revealed that a subtle rotation of the plane of the scissile bond allows the carbonyl carbon to be attacked from alternate directions in the two steps of the reaction, and the location of the catalytic His residue allows it to act as a proton acceptor or donor to the nucleophile or leaving group, respectively, with little or no change in position. Comparison of the structure of the complex with those representing other intermediates published previously, demonstrated that the residues of the catalytic triad are positioned to promote each step of both the forward and reverse reactions with remarkably little motion and with conservation of hydrogen bonding interactions. The results provide insights into the mechanism by which BPTI and possibly cyclic peptide inhibitors such as the brunsvicamides resist hydrolysis when bound to their target proteases.

Another interesting aspect in resisting enzymatic hydrolysis of cyclic peptide inhibitors by the respective protease is the prevention of access of water to the catalytic site thus avoiding deacylation. The crystal structure of the elastase-scyptolin A complex showed that the macrocycle occupies a crucial part of the active site thereby preventing the access of hydrolytic water and thus cleavage (McDonough and Schofield, 2003). Interestingly, the distance between the carbonyl carbon of the scissile bond and the oxygen atom of the catalytic serine residue was found to be within interaction distances. This suggests that the acylation step indeed occurs but due to prevention of the access of a water

molecule that is required for the second step in the hydrolysis, the deacylation step is prevented.

Conformational constraints also contribute to proteolytic stability of the protease-bound inhibitor. Proteases generally recognize their substrates and inhibitors in an extended  $\beta$ -strand conformation (Tyndall et al., 2005). Some macrocyclic peptidomimetics are constrained mimics of linear peptides with a preorganized  $\beta$ -strand conformation for protease binding. This concept has been implemented in the synthesis of macrocyclic protease inhibitors designed to be constrained into a  $\beta$ -strand-like geometry (Abell et al., 2009). It has also been shown that certain macrocyclic natural products present a short, extended  $\beta$ -strand to proteases (Loughlin et al., 2004). For example, the 19-membered thrombin inhibitor cyclotheonamide A adopts an extended  $\beta$ -strand conformation of its protease-binding region (D-Phe-Arg-Pro) (Greco et al., 1996). In the elastase-scyptolin A complex, four N-terminal amino acids (Leu-Thr-Thr-Ala) bind at subsites S1 through S4 of elastase forming hydrogen bonds, similar to those found in an antiparallel  $\beta$ -sheets (Matern et al., 2003). The corresponding substructure of brunsvicamide B (Ile-Lysurea-Ile) has the potential to form similar hydrogen-bonding interactions inside the active site of HLE. However, deviations from an extended  $\beta$ -strand conformation within the macrocyclic structure, as can be observed for both scyptolin A and brunsvicamide B, might result in repositioning of the cleavage site relative to the catalytic triad. Future investigations are needed to clarify whether scyptolin A and brunsvicamide B are cleaved and recycled within the active site of HLE.

The putative binding mode of brunsvicamides in the active site of HLE suggests that their inhibitory profile might be largely governed by the residue Val (or Ile) occupying the P1 position. Brunsvicamides inhibit HLE, an enzyme that has a primary specificity for small hydrophobic residues, but not chymotrypsin, cathepsin G or trypsin, which prefer an aromatic or basic moiety, respectively, at the P1 position. Related cyclic cyanopeptides containing an exocyclic urea moiety at the N-terminal D-Lys are anticipated to bind in a similar manner to elastases, and the generated brunsvicamide-HLE model could be used to explain the inhibitory profile of these cyclic peptides with respect to their structure. For example, oscillamide Y (Marsh et al., 1997) or anabaenopeptins B, E, F, and T (Itou et al., 1999; Kodani et al., 1999; Bubik et al., 2008) with Ile/Val in the C-terminal position next to D-Lys (i.e. P1) do not inhibit chymotrypsin.

On the other hand, the inhibitory activity of anabaenopeptins against carboxypeptidase A is obviously caused by a different mode of interaction. Their potency mainly depends on the exocyclic amino acid attached to the ureido group. Carboxypeptidase A is an exopeptidase with a substrate

preference for hydrophobic amino acids at the C-terminal position, which bind to the S1' pocket (Vendrell, et al., 2000). Those anabaenopeptins with a terminal basic amino acid at the ureido group (B, E, F, H) showed lower potency than those with a hydrophobic residue (G, I, J, T), (Itou et al., 1999; Kodani et al., 1999; Murakami et al., 2000) indicating that this residue occupies the S1' pocket of carboxypeptidase A.

## 6.4 Summary

In summary, the three cyclic cyanobacterial peptides were found to be selective inhibitors of HLE. They did not inhibit the related serine proteases, serine esterases and cysteine proteases. Molecular modeling studies showed that the brunsvicamides mimic the binding mode of the experimentally determined scyptolin A in complex with PPE. The mechanism by which the cyclic peptides resist proteolytic cleavage and remain tightly bound to the active site is probably due to the equilibrium between cleavage and reformation of the peptide bond. The constrained conformation of the cyclic peptides also contributes to the observed inhibition. Finally, the brunsvicamides investigated here can be used as potential leads to further design a highly potent and stable inhibitor of HLE for use in the treatment of emphysema.



# Chapter 7

## Summary and conclusions

This chapter presents an overall summary of the findings of the thesis. The major objectives of the thesis were VS method development and practical applications for identification of inhibitors of selected pharmaceutical targets, two cysteine proteases and a membrane-bound serine protease. Additional enzyme inhibition and molecular modeling studies on three cyclic peptides against human leukocyte elastase were also performed.

In the first part of the thesis, three-dimensional protein-ligand interaction information was successfully applied for the development of a new hybrid VS method. The methodology, termed the interaction annotated structural features (IASF), was introduced that assigns energy-based scores to two-dimensional substructures based on three-dimensional protein-ligand interaction information extracted by using a scoring function. The performance of the new method was evaluated in real HTS screening sets and was found to perform better than conventional fragment-based 2D fingerprint similarity searching and three-dimensional docking calculations. The performance results indicate the information gain in 2D substructure searching when 3D interaction information is integrated.

The second aim of the thesis was analysis of the nature of SARs in analogue series at molecular 3D protein-ligand interaction level. Different compound series in combinatorial analog graphs were analyzed and substitution patterns that introduce activity cliffs of varying magnitude were determined. The systematically identified SAR determinants were then studied on the basis

of three-dimensional ligand-target interaction to enable a structural interpretation of SAR discontinuity and underlying activity cliffs. The results showed that many discontinuous SAR features extracted from combinatorial analog graphs can be directly associated with experimental three-dimensional receptor-ligand interactions. However, some substitution site patterns that introduce significant SAR discontinuity in analog series were not clearly explainable based on only protein-ligand interaction information.

In the second part of the thesis, as a practical experiment of applications of VS methods, different approaches were implemented for the identification of selected cysteine (cathepsin K and S) and a membrane-bound serine protease (matriptase-2) inhibitors. By applying the compound mapping algorithm, DynaMAD, from a database containing ~3.7 million compounds 10 candidate compounds were selected and tested. This resulted in the identification of two dual inhibitors of cathepsin K and S with new scaffolds. Both the identified inhibitors did not contain an electrophilic “warhead” that usually is present in most of the previously reported covalently interacting cathepsin inhibitors.

In a similar VS application, through combined SB and LB approach supported by knowledge-based compound design, two *N*-protected dipeptide amides containing a 4-amidinobenzylamide were identified as the first small molecule inhibitors of matriptase-2 with  $K_i$  values of 170 nM and 460 nM, respectively. A new inhibitor of the closely related protease, matriptase-1, was also identified with a  $K_i$  value of 220 nM showing more than 50-fold selectivity over matriptase-2.

Finally, three cyclic cyanobacterial peptides, brunsvicamides A-C, were tested against HLE and a panel of other serine proteases, serine esterases and a cysteine protease. The peptides were found to be potent and selective inhibitors of HLE. Molecular modeling studies were performed to get an insight into their possible binding mode. The results showed that the brunsvicamides form several favorable intermolecular interaction and they mimic the binding mode of the experimentally determined scryptolin A in complex with porcine pancreatic elastase.

These newly identified VS hits and the HLE inhibiting cyclic peptides provide starting points for further chemical exploration of new potential inhibitors of the respective enzymes.

# Appendices

## A. Software and Databases

Software and databases used for the various studies presented in this thesis are given in the following tables and are ordered alphabetically.

<b>BindigDB</b>	Skaggs School of Pharmacy and Pharmaceutical Sciences, California, USA
Description	BindingDB is a public database organized around the concept of the binding assay. It contains data on measured binding affinities of small drug-like molecules against relevant drug targets (Liu et al., 2007).
WebSite	<a href="http://www.bindingdb.org/">http://www.bindingdb.org/</a>

<b>DOCK6</b>	University of California, San Francisco, USA
Description	DOCK6 is a suite of automated molecular docking tools designed to predict binding modes of small molecules to a protein target (Meng et al., 1992).
WebSite	<a href="http://dock.compbio.ucsf.edu/">http://dock.compbio.ucsf.edu/</a>

<b>DynaMAD</b>	Life Science Informatics, University of Bonn, Germany
Description	DynaMAD is designed to map database compounds to activity-specific consensus positions in chemical space representations of step-wise increasing dimensionality (Eckert et al., 2006).
WebSite	<a href="http://www.lifescienceinformatics.uni-bonn.de/">http://www.lifescienceinformatics.uni-bonn.de/</a>

<b>FlexX</b>	BioSolveIT GmbH, Sankt Augustin, Germany
Description	FlexX is a fast flexible docking tool using an incremental construction algorithm that first places a base fragment in the active site and then extends it to peripheral fragments according to the most favorable torsion and protein-ligand interactions (Rarey et al., 1996).
WebSite	<a href="http://www.biosolveit.de/">http://www.biosolveit.de/</a>

<b>MACCS</b>	Symyx Software, San Ramon, CA, USA
Description	MACCS structural keys represent a 2D fingerprint that consists of 166 bits coding for 166 structural fragments (McGregor and Pallai, 1997).
WebSite	<a href="http://www.mdl.com/">http://www.mdl.com/</a>

<b>MDDR</b>	MDL Information Systems Inc., San Leandro, USA
Description	MDDR is a database that contains about 160,000 therapeutically (targetwise) annotated biologically active compounds.
WebSite	<a href="http://www.mdl.com/products/knowledge/drug_data_report/">http://www.mdl.com/products/knowledge/drug_data_report/</a>

<b>MOE</b>	Chemical Computing Group Inc., Montreal, Canada
Description	The Molecular Operating Environment (MOE) is a suit of molecular modeling tools which provides applications for computational modeling works.
WebSite	<a href="http://www.chemcomp.com/">http://www.chemcomp.com/</a>



<b>Molprint2D</b>	Unilever Centre for Molecular Science Informatics, Cambridge, UK
Description	Molprint2D represents layered atom environment-based 2D structural fingerprints of a molecule (Bender et al., 2004).
WebSite	<a href="http://www.molprint.com/">http://www.molprint.com/</a>

<b>PDBbind</b>	Shaomeng Wang Laboratory, University of Michigan, USA
Description	PDBbind is a comprehensive collection of experimentally measured binding affinity data for protein-ligand complexes deposited in the PDB (Wang et al., 2004).
WebSite	<a href="http://www.pdbbind.org/">http://www.pdbbind.org/</a>

<b>PipeLinePilot</b>	Accelrys Inc., San Diego, USA
Description	Scitegic Pipeline Pilot is a graphical software for creating workflow protocols and provides components for data analysis and various scientific applications.
WebSite	<a href="http://www.accelrys.com/products/scitegic/">http://www.accelrys.com/products/scitegic/</a>

<b>PubChem</b>	National Center for Biotechnology Information, MD, USA
Description	PubChem is a comprehensive public database that provides information on the biological activities of small molecules.
WebSite	<a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a>

<b>PyMOL</b>	DeLano Scientific LLC, California, USA
Description	PyMOL is a free open-source molecular visualization tool.
WebSite	<a href="http://www.pymol.org/">http://www.pymol.org/</a>

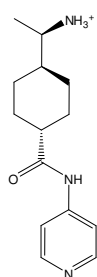
<b>Python</b>	The Python Software Foundation
Description	Python is a dynamic programming language that is used in a wide variety of application domains.
WebSite	<i><a href="http://www.python.org/">http://www.python.org/</a></i>

<b>ZINC</b>	University of California, San Francisco, USA
Description	ZINC is a free database of commercially available compounds in predicted 3D conformational states (Irwin and Shoichet, 2005).
WebSite	<i><a href="http://blaster.docking.org/zinc/">http://blaster.docking.org/zinc/</a></i>

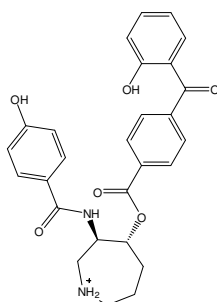
## B. Reference ligands from complex crystal structures

The following tables present 2D structures of X-ray reference crystal ligands used in *Chapter 2*. Below each structure, the corresponding PDB ID is given.

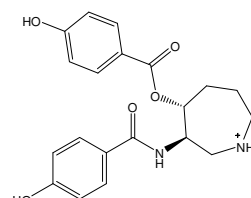
**Figure B.1: Protein Kinase A (PKA).**



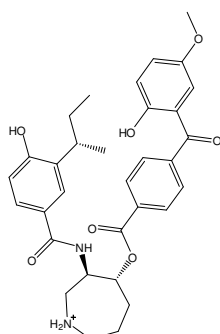
1Q8T



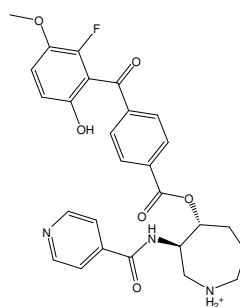
1RE8



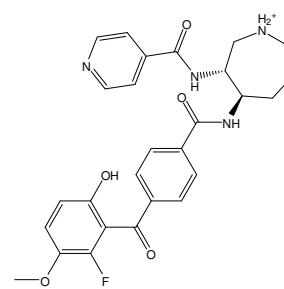
1REJ



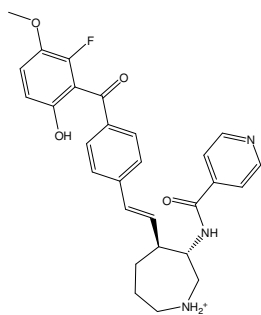
1REK



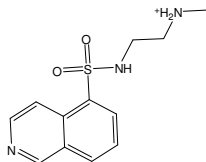
1SVE



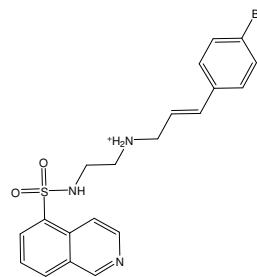
1SVG

**Figure B.1: (continued) Protein Kinase A (PKA).**

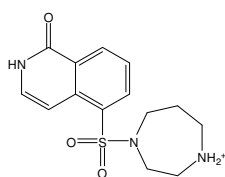
1SVH



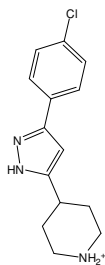
1YDS



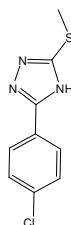
1YDT



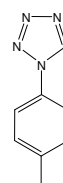
2ERZ

**Figure B.2: Thrombin (THR).**

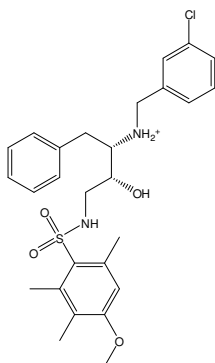
1WAY



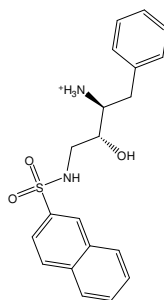
1WBG



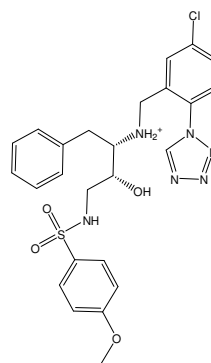
2C90



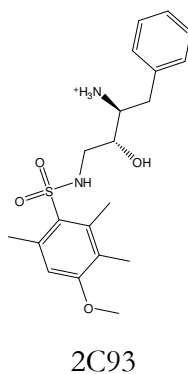
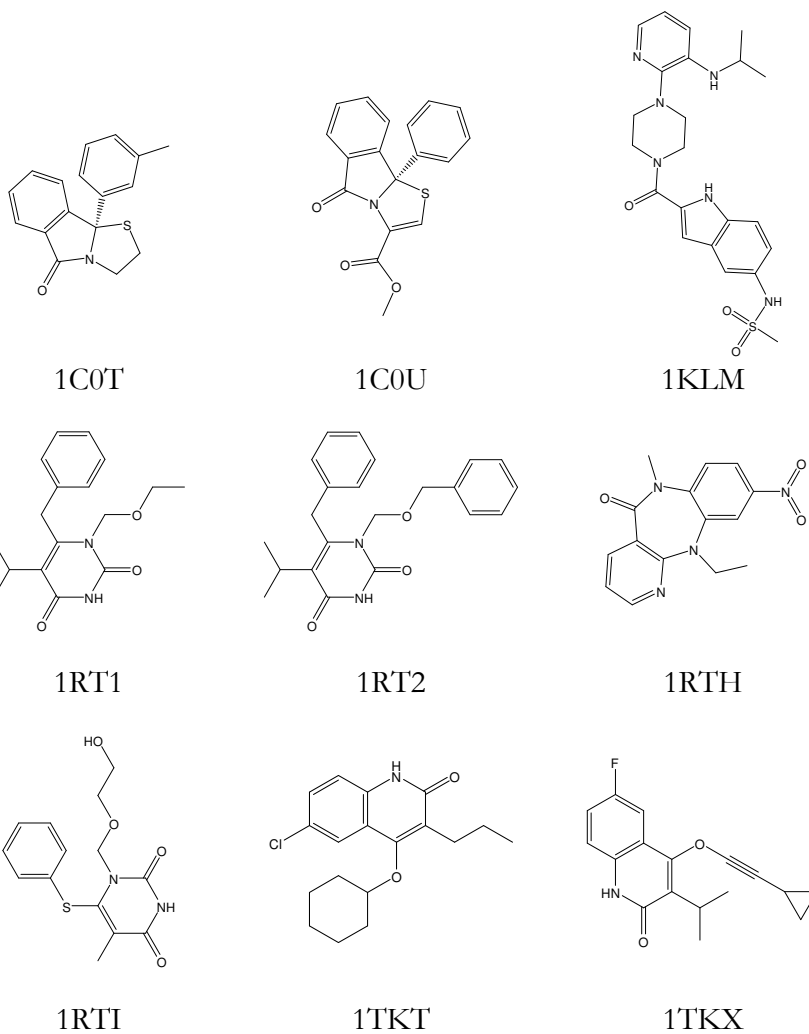
2C8X



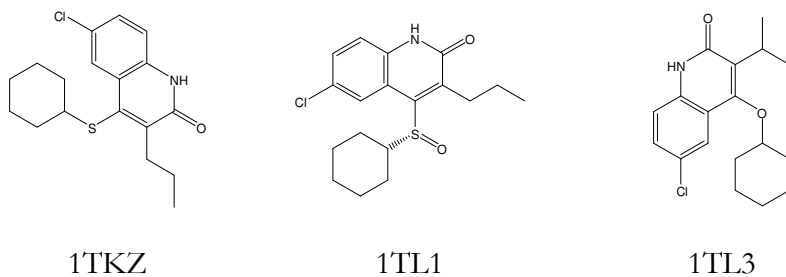
2C8Y



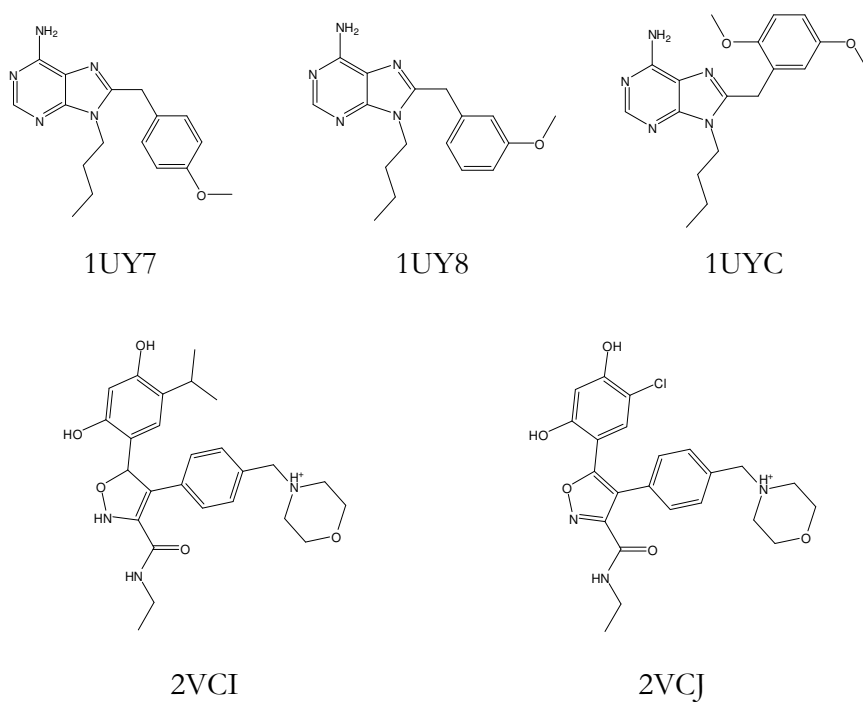
2C8W

**Figure B.2: (continued) Thrombin (THR).****Figure B.3: Human immunodeficiency virus reverse transcriptase (HIV).**

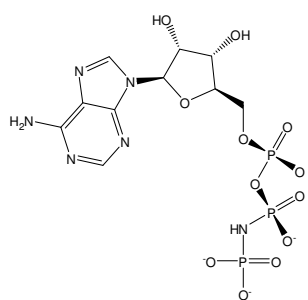
**Figure B.3: (continued) Human immunodeficiency virus reverse transcriptase (HIV).**



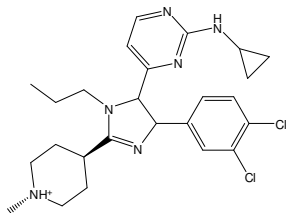
**Figure B.4: Heat shock protein 90 (HSP).**



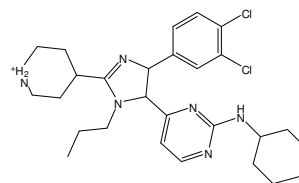
**Figure B.5: c-jun N-terminal kinase 3 (JNK).**



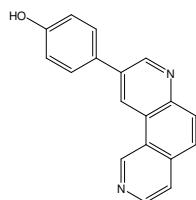
1JNK



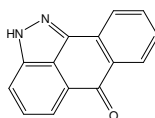
1PMN



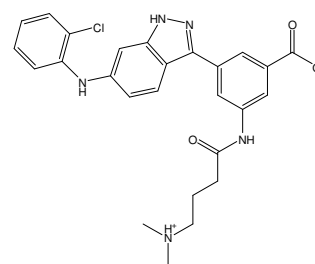
1PMQ



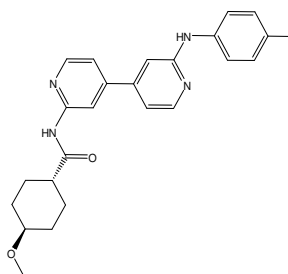
1PMU



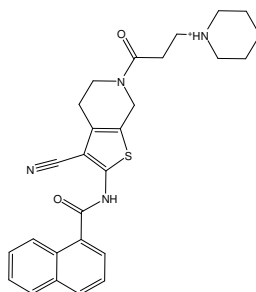
1PMV



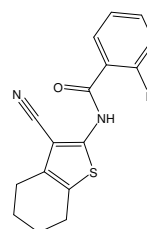
2B1P



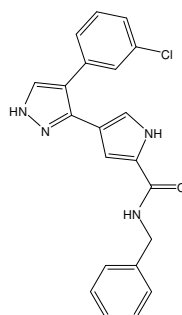
2EXC



2O0U



2O2U



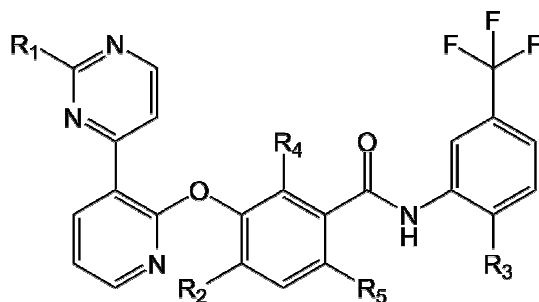
2OK1

## C. SAR tables

The SAR tables here present the individual compounds, corresponding R-groups ('R1', 'R2', ...) and potency values for all compounds in the four analog series (Tie-2 kinase, Factor Xa (series 1 and 2) and thrombin) discussed in *Chapter 3*. Compounds are identified by the original BindingDB monomer id (except for thrombin where the original publication numbering is used). The core structure is presented on top of each table and attachment points are marked with the letter 'Z'. For the reference X-ray ligand, the PDB ligand identifier is given in parentheses.

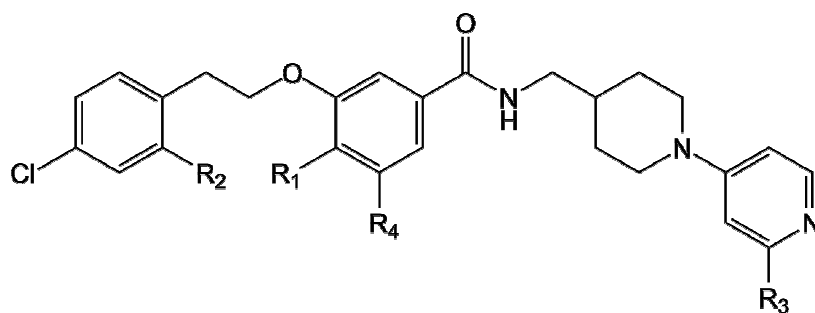


Table C.1: Tie-2 kinase inhibitors.

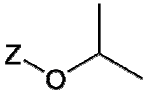


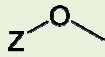
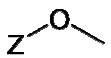

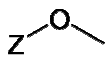
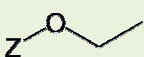

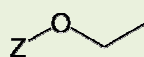
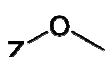




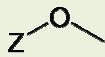
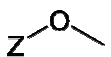

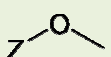


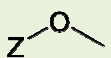


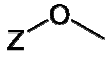




BindingDB monomer id	Potency [nM]	R1	R2	R3	R4	R5
14948	1					
14977	153					
14982	399					
14983 (MR9)	10					
14989	99					
14992	39					
14993	388					
14995	4					

Table C.2: Factor Xa inhibitors (series 1).

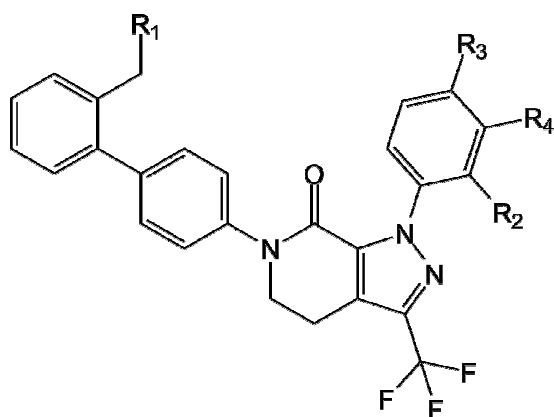


BindingDB monomer id	Potency [nM]	R1	R2	R3	R4
13616	50				
13639	106				
13641	156				
13642	37				
13645	29				
13646	48				
13650	588				
13651	51				
13653	61				


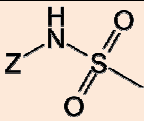

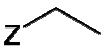

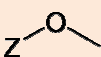
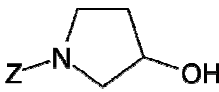
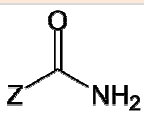
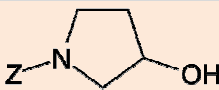
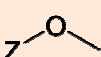
BindingDB monomer id	Potency [nM]	R1	R2	R3	R4
13654	233				
13655	125				
13656	50				
13657	263				
13661	1053				
13663	13				
13664 (I1H)	18				
13665	41				
13667	41				
13673	57				
13678	950				

**Table C.2:** (continued) Factor Xa inhibitors (series 1).

Table C.3: Factor Xa inhibitors (series 2).

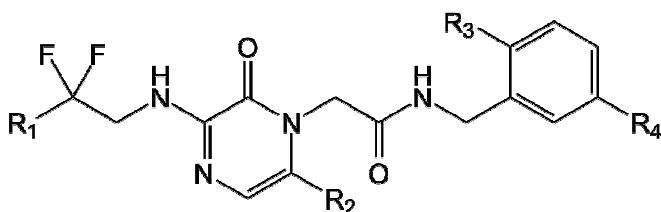


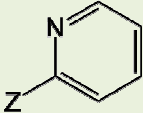

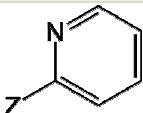


BindingDB monomer id	Potency [nM]	R1	R2	R3	R4
12730	0.82				
12731	2.7				
12732	12				
12733 (4QC)	0.18				
12734	20				
12735	2				
12736	88				
12737	54				

BindingDB monomer id	Potency [nM]	R1	R2	R3	R4
12738	47				
12739	2.6				
12740	0.35				
12742	0.72				
12743	0.18				

**Table C.3:** (continued) Factor Xa inhibitors (series 2).

**Table C.4:** Thrombin inhibitors.



Compound	Potency [nM]	R1	R2	R3	R4
4	12				
5	0.44				

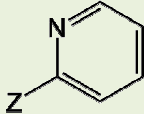

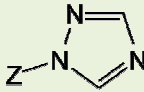
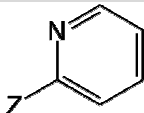

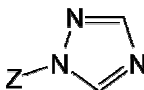

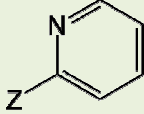

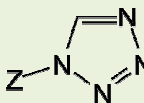
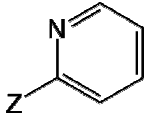

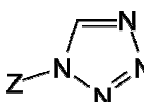

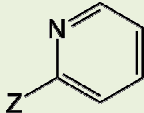

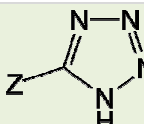
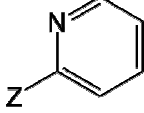
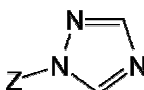
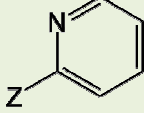
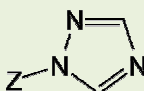

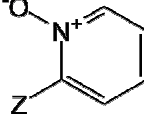

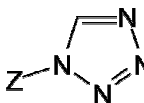
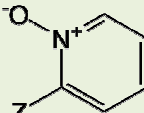

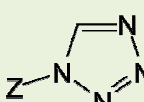

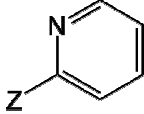
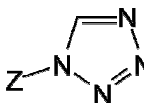
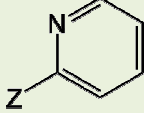
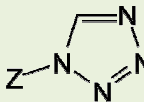

Compound	Potency [nM]	R1	R2	R3	R4
13	0.45				
14	0.01				
16	0.1				
17	0.0015				
19	940				
24	16				
25	0.24				
33	0.05				
34 (170)	0.0014				
35	2.7				
36	0.033				

Table C.4: (continued) Thrombin inhibitors.

## D. Screening data sets

### Active reference data source

The following table shows summary of the sources of the 42 reference compounds used in cathepsin K and S inhibitor searching (*chapter 4*). The numbering of the compounds in the original scientific literature is given for reference.

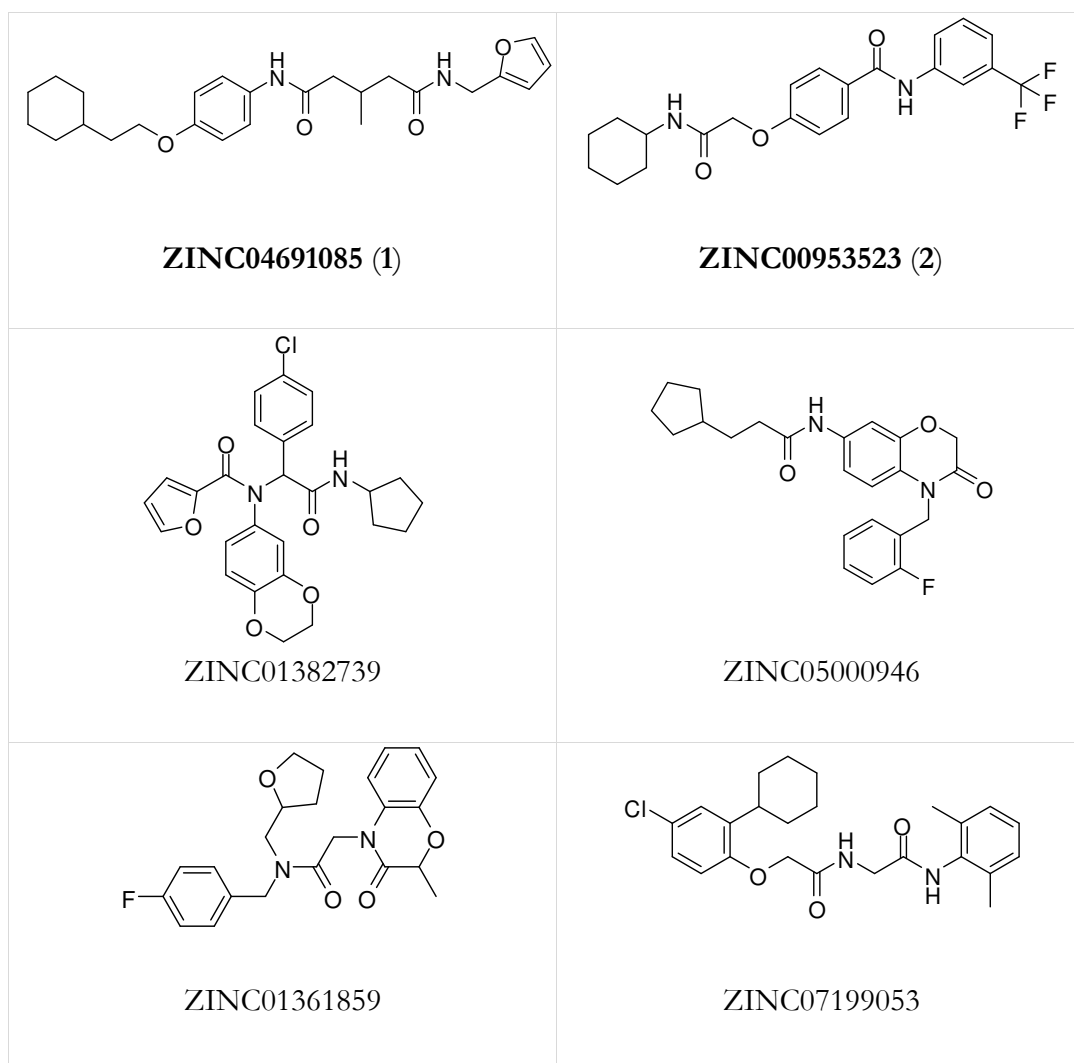
**Table D.1. Reference compounds used in cathepsin K and S inhibitor searching.**

Compound	Original reference
21	(Chatterjee et al., 2007)
12d	(Inagaki et al., 2007)
JNJ	(Thurmond et al., 2004a)
123689	MDDR
15, 19, 20	(Lui et al., 2005)
3d, 3e, 3f	(Thurmond et al., 2004b)
11c, 11d, 11e	(Patterson et al., 2006)
14, 15, 18, 23	(Tully et al., 2006b)
3, 9, 10, 11, 12, 13, 14, 17, 18, 19, 20	(Tully et al., 2006a)
6, 9, 10, 11, 12, 13, 14, 15, 17, 19, 20, 21, 24, 25	(Gauthier et al., 2007)

## Selected candidate compounds

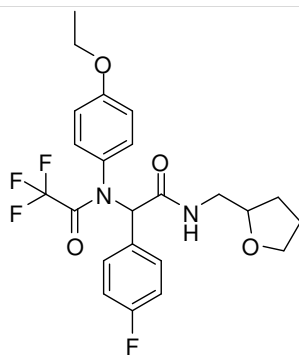
The following figure shows the 2D structures of the candidate compounds selected from the virtual screening search calculations in *Chapter 4*.

**Figure D.1: Structures of the 10 tested candidate molecules from virtual screening of cathepsin K and S inhibitors.** ZINC IDs of the two dual cathepsin K and S inhibitors identified are printed in bold and the corresponding numbers are given in brackets.

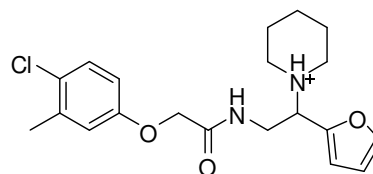




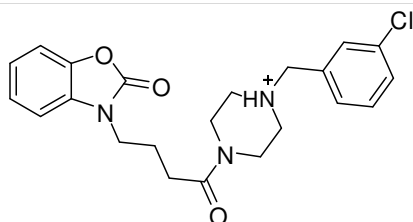
**Figure D.1: (continued) Structures of the 10 tested candidate molecules from virtual screening of cathepsin K and S inhibitors. ZINC IDs of the two dual cathepsin K and S inhibitors identified are printed in bold.**



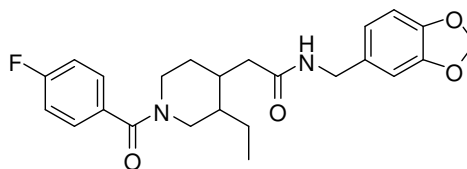
ZINC06600967



ZINC07406375



ZINC07544113



ZINC08296007

## Sources of compounds

**Table D.2: Suppliers from which the candidate molecules were purchased.** ZINC IDs and corresponding supplier catalog (order) numbers are given. The compounds identified as active are indicated in bold.

ZINC ID	Supplier	Catalog number
<b>ZINC04691085</b>	<b>Vitas-M</b>	<b>STK117823</b>
<b>ZINC00953523</b>	<b>Life Chemicals</b>	<b>F3098-6036</b>
ZINC01382739	Asinex	ASN05255922
ZINC01361859	Asinex	ASN06912209
ZINC06600967	Asinex	BAS06264611
ZINC05000946	Asinex	ASN10030324
ZINC07544113	Enamine-REAL	Z52266270
ZINC07199053	Enamine-REAL	ZU-4220039
ZINC07406375	Enamine-REAL	ZU-1831050
ZINC08296007	AnalytiCon	NAT14-316439

**Table D.3: ZINC IDs of commercially unavailable compounds.** The ZINC IDs of candidate compounds selected from virtual screening runs which were not available for purchase, described in *chapters 4 and 5*, are given.

Screening system	ZINC IDs	
Cathepsin K and S	ZINC03039475	ZINC04547528
	ZINC07447875	ZINC02016642
	ZINC03993509	ZINC04560102
	ZINC03963240	ZINC04056841
	ZINC07850162	ZINC03827073
	ZINC04505181	ZINC06716127
Matriptase-2	ZINC03838230	ZINC01549765
	ZINC03808361	ZINC01549825
	ZINC03966454	ZINC03818000
	ZINC01996748	ZINC03921255
	ZINC01548696	ZINC03808363
	ZINC03831686	

## D. Laboratory experimental details

The following sections provide detailed laboratory experimental procedures for compounds discussed in chapters 4, 5 and 6.

### Cathepsin K inhibition assay

A fluorometric assay (Perkin Elmer luminescence spectrometer LS 55) was used to measure the activity of recombinant human cathepsin K (expressed in *Pichia pastoris*) at 22 °C. The wavelengths for excitation and emission were 360 nm and 490 nm, respectively. Assay buffer was 100 mM sodium citrate buffer, pH 5.0, 100 mM NaCl, 1 mM EDTA, 0.01% CHAPS. An enzyme stock solution of 1.8  $\mu$ M in assay buffer was diluted 1:100 with assay buffer containing 5 mM DTT and incubated for 30 min at 37 °C. Inhibitor stock solutions were prepared in DMSO. A 20 mM stock solution of the chromogenic substrate Z-Leu-Arg-NH-Mec (Bachem, Bubendorf, Switzerland) was prepared with DMSO. The final concentration of DMSO was 5% and the final concentration of the substrate Z-Leu-Arg-NH-Mec was 40  $\mu$ M. Assays were performed with a final concentration of 0.18 nM of cathepsin K. Into a cuvette containing 940  $\mu$ L assay buffer, inhibitor solution and DMSO in a total volume of 48  $\mu$ L, and 2  $\mu$ L of the substrate solution were added and thoroughly mixed. The reaction was initiated by adding 10  $\mu$ L of the cathepsin K solution and was followed over 8 min. IC<sub>50</sub> values were calculated from the linear steady-state turnover of the substrate. A  $K_m$  value of  $5.8 \pm 0.4$   $\mu$ M was obtained in duplicate measurements with eight different substrate concentrations. Inhibitory activity, expressed as IC<sub>50</sub> value, was determined from the linear steady-state turnover of the substrate in triplicate measurements at a single inhibitor concentration.

## Cathepsin S inhibition assay

Recombinant human cathepsin S (Calbiochem, Darmstadt, Germany) was assayed spectrophotometrically (Cary 100 Bio, Varian) at 405 nm at 37 °C. Assay buffer was 50 mM sodium phosphate buffer, pH 6.5, 50 mM NaCl, 2 mM EDTA, 0.01% Triton X-100. An enzyme stock solution of 866 µg/mL in 35 mM potassium phosphate, 35 mM sodium acetate, 2 mM DTT, 2 mM EDTA, 50% ethylene glycol, pH 6.5. This enzyme solution was diluted 1:100 with assay buffer containing 5 mM DTT and incubated for 30 min at 37 °C. Inhibitor stock solutions were prepared in DMSO. A 10 mM stock solution of the chromogenic substrate Z-Phe-Val-Arg-pNA (Bachem, Bubendorf, Switzerland) was prepared with DMSO. The final concentration of DMSO was 5% and the final concentration of the substrate Z-Phe-Val-Arg-pNA was 100 µM. Assays were performed with a final concentration of 86.6 ng/mL of cathepsin S, which corresponded to an initial rate of 0.6 µM/min. Into a cuvette containing 940 µL assay buffer, inhibitor solution and DMSO in a total volume of 40 µL, and 10 µL of the substrate solution were added and thoroughly mixed. The reaction was initiated by adding 10 µL of the cathepsin S solution and was followed over 18 min. IC<sub>50</sub> values were calculated from the linear steady-state turnover of the substrate. A  $K_m$  value of  $75 \pm 7$  µM was obtained in duplicate measurements with nine different substrate concentrations. Inhibitory activity, expressed as IC<sub>50</sub> value, was determined from the linear steady-state turnover of the substrate in triplicate measurements at a single inhibitor concentration.

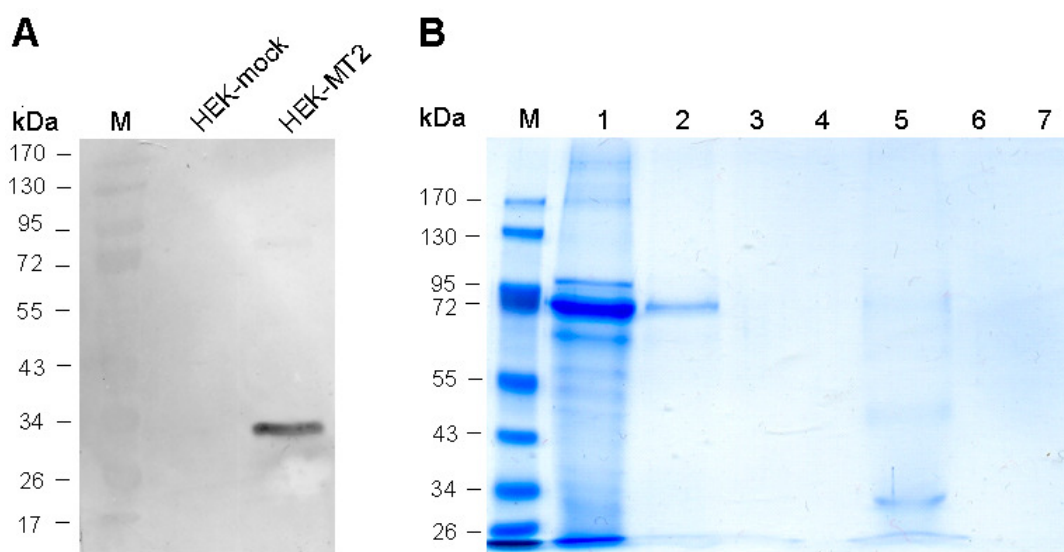
## Cathepsin L inhibition assay

Recombinant human cathepsin L (Calbiochem, Darmstadt, Germany) was assayed spectrophotometrically (Cary 100 Bio, Varian) at 405 nm at 37 °C. Assay buffer was 100 mM sodium phosphate buffer, pH 6.0, 100 mM NaCl, 5 mM EDTA, 0.01% Brij 35. An enzyme stock solution of 50 µg/mL in 20 mM sodium acetate buffer, pH 5.0, 100 mM NaCl, 10 mM trehalose, 1 mM EDTA, 50% glycerol was diluted 1:100 with assay buffer containing 5 mM DTT and incubated for 30 min at 37 °C. This enzyme solution was diluted 1:5 with assay buffer containing 5 mM DTT. Inhibitor stock solutions were prepared in DMSO. A 10 mM stock solution of the chromogenic substrate Z-Phe-Arg-pNA (Bachem, Bubendorf, Switzerland) was prepared with DMSO. The final concentration of DMSO was 5% and the final concentration of the substrate Z-Phe-Arg-pNA was 100 µM. Assays were performed with a final concentration of 4 ng/mL of cathepsin L, which corresponded to an initial rate of 0.9 µM/min. Into a cuvette containing 910 µL assay buffer, inhibitor solution and DMSO in a total volume of 40 µL, and 10 µL of the substrate solution were

added and thoroughly mixed. The reaction was initiated by adding 40  $\mu\text{L}$  of the cathepsin L solution and was followed over 10 min.  $\text{IC}_{50}$  values were calculated from the linear steady-state turnover of the substrate. A  $K_m$  value of  $16 \pm 1 \mu\text{M}$  was obtained in duplicate measurements with nine different substrate concentrations. Inhibitory activity, expressed as  $\text{IC}_{50}$  value, was determined from the linear steady-state turnover of the substrate in triplicate measurements at a single inhibitor concentration.

## Matriptase-2 expression and purification

The catalytic domain of recombinant human matriptase-2 is not commercially available and therefore, to study the inhibitory effect of compounds **1-4** (discussed in *chapter 5*) on human matriptase-2, the whole matriptase-2 construct was cloned and expressed in HEK (human embryonic kidney) cells with a Myc-tag at the C-terminal end of the protein. As shown in Figure E.1a an approximately 30 kDa C-terminal fragment of matriptase-2 was detectable using anti-c-Myc antibody in the conditioned medium of transfected HEK cells (HEK-MT2). This form represents the catalytic domain released from the cell surface after processing of the zymogen, as described previously (Silvestri et al., 2008). No signal was detectable in HEK cells expressing the empty vector (HEK-mock). The catalytic domain of matriptase-2 was purified from the conditioned medium of HEK-MT2 cells by immunoaffinity chromatography. After isolation, the purity of matriptase-2 was checked on SDS-Page (Figure E.1b). One single band of approximately 30 kDa was visible after purification representing the catalytic domain of matriptase-2.



**Figure E.1: Expression and purification of human matriptase-2.** (A) Conditioned medium of HEK cells transfected with an expressing vector harbouring matriptase-2 cDNA (HEK-MT2) and without matriptase-2 cDNA (HEK-mock) was characterized by western blot analysis using anti-c-Myc antibody. (B) The catalytic domain of matriptase-2 was isolated by affinity chromatography using immobilized anti-c-Myc antibody. Lane 1: flow-through; lanes 2-4: wash fractions; lanes 5-7: elution fractions; M: molecular mass markers.

The activity profile of matriptase-2 in the conditioned medium and of the purified catalytic domain of matriptase-2 was characterized using the chromogenic substrate Boc-Gln-Ala-Arg-*para*-nitroanilide. Similar  $K_m$  and  $K_i$  values were obtained from the experiments with either conditioned medium or purified enzyme. Moreover, no detectable activity was observed in the conditioned medium of non-transfected HEK cells or HEK cells expressing the empty vector (HEK-mock). Thus, it is possible to use the conditioned medium as a source for matriptase-2 activity. The  $K_m$  values for matriptase-2 in the conditioned medium (210  $\mu\text{M}$ ) and for purified matriptase-2 (159  $\mu\text{M}$ ) in the micromolar range were similar to human recombinant matriptase-1 (refer to Table 5.2) and to the reported value (257  $\mu\text{M}$ ) (Cho et al., 2001) for the purified catalytic domain of murine matriptase-1 overexpressed in insect cells measured with the corresponding fluorogenic substrate.

## Matriptase-2 and matriptase-1 inhibition assays

The activity of matriptase-2 in the conditioned medium of HEK-MT2 cells, of the purified catalytic domain of matriptase-2 and of recombinant matriptase-1 (catalytic domain; Enzo Life Sciences, Lörrach, Germany) was assayed in Tris saline buffer (50 mM Tris, 150 mM NaCl, pH 8.0) at 37°C by monitoring the release of *para*-nitroaniline from the chromogenic substrate Boc-Gln-Ala-Arg-

*para*-nitroanilide (Bachem, Bubendorf, Switzerland) at 405 nm using a Cary 100 UV-Vis spectrophotometer (Varian, Darmstadt, Germany).  $K_m$  values were determined with eight different substrate concentrations in duplicate experiments. Inhibition assays were performed in duplicate or triplicate measurements with three (**2** and **4** at matriptase-2), or at least five (other experiments) different inhibitor concentrations.  $IC_{50}$  values were obtained by non-linear regression according to the equation  $v = v_0 / (1 + [I]/IC_{50})$ . 10 mM inhibitor stock solutions and a 100 mM stock solution of Boc-Gln-Ala-Arg-*para*-nitroanilide were prepared in DMSO. The final concentration of the substrate was 400  $\mu$ M and of DMSO was 1.5%. Into a cuvette containing 979  $\mu$ L pre-warmed assay buffer, 11  $\mu$ L of an inhibitor solution and 4  $\mu$ L of a substrate solution were added and thoroughly mixed. The reaction was initiated by adding 6  $\mu$ L of an enzyme solution (5  $\mu$ g / 6  $\mu$ L total protein of the conditioned medium of HEK-MT2 cells; 28 ng / 6  $\mu$ L purified catalytic domain of matriptase-2; 3 ng / 6  $\mu$ L of matriptase-1) and was followed over 20 min.

## HLE inhibition assay

Human leukocyte elastase (Calbiochem, Darmstadt, Germany) was assayed spectrophotometrically (Varian, Cary 50 Bio) at 405 nm at 25 °C for 10 min. Inhibitor stock solutions (brunsvicamides A-C were isolated as described by Müller et al. (Müller et al., 2006) and were provided by Prof. Dr. G. König and Dr. C. Mehner) were prepared in DMSO. Assay buffer was 50 mM sodium phosphate buffer, 500 mM NaCl, pH 7.8. An enzyme stock solution of 50  $\mu$ g/mL was prepared in 100 mM sodium acetate buffer, pH 5.5 and diluted with assay buffer. A 50 mM stock solution of the chromogenic substrate MeOSuc-Ala-Ala-Pro-Val-pNA (Bachem, Bubendorf, Switzerland) was prepared in DMSO and diluted with assay buffer containing 10% DMSO. The final concentration of the substrate was 100  $\mu$ M, of DMSO was 1.5% and of HLE was 50 ng/mL. Into a cuvette containing 890  $\mu$ L assay buffer, 10  $\mu$ L of an inhibitor solution and 50  $\mu$ L of a substrate solution were added and thoroughly mixed. The reaction was initiated by adding 50  $\mu$ L of the HLE solution. A three-parameter model  $v = v_0 / (1 + ([I]/IC_{50})^x)$  was used for non-linear regression.

## HLE incubation experiments

To 450  $\mu\text{L}$  of assay buffer, 10  $\mu\text{L}$  of an HLE (Calbiochem, Darmstadt, Germany) solution (50  $\mu\text{g}/\text{mL}$ ) and 40  $\mu\text{L}$  of a solution of brunsvicamide C (1.25  $\text{mg}/\text{mL}$  in DMSO) or 40  $\mu\text{L}$  DMSO were added. The mixtures were incubated at 25  $^{\circ}\text{C}$  and aliquots of 50  $\mu\text{L}$  were added to a cuvette containing 894  $\mu\text{L}$  of assay buffer and 50  $\mu\text{L}$  of a solution of MeOSuc-Ala-Ala-Pro-Val-pNA (2  $\text{mM}$  in assay buffer with 10% DMSO) and 6  $\mu\text{L}$  DMSO. Final concentrations were as follows: 50  $\text{ng}/\text{mL}$  HLE, 100  $\mu\text{M}$  MeOSuc-Ala-Ala-Pro-Val-pNA, 5  $\mu\text{g}/\text{mL}$  brunsvicamide C and 1.5% DMSO. Reactions were followed at 25  $^{\circ}\text{C}$  for 10 min at 405 nm.

## Cathepsin G inhibition assay

Human cathepsin G (Calbiochem, Darmstadt, Germany) was assayed spectrophotometrically (Varian, Cary 50 Bio) at 405 nm at 25  $^{\circ}\text{C}$  for 10 min. Inhibitor stock solutions were prepared in DMSO. Assay buffer was 20  $\text{mM}$  Tris HCl buffer, 150  $\text{mM}$  NaCl, pH 8.4. An enzyme stock solution of 200  $\text{mU}/\text{mL}$  was prepared in 50  $\text{mM}$  sodium acetate buffer, 150  $\text{mM}$  NaCl, pH 5.5. A 50  $\text{mM}$  stock solution of the chromogenic substrate Suc-Ala-Ala-Pro-Phe-pNA (Bachem, Bubendorf, Switzerland) in DMSO was diluted with assay buffer. The final concentration of the substrate was 500  $\mu\text{M}$ , of DMSO was 1.5% and of cathepsin G was 2.5  $\text{mU}/\text{mL}$ . Into a cuvette containing 882.5  $\mu\text{L}$  assay buffer, 5  $\mu\text{L}$  of an inhibitor solution and 100  $\mu\text{L}$  of a substrate solution were added and thoroughly mixed. The reaction was initiated by adding 12.5  $\mu\text{L}$  of the cathepsin G solution.

## Chymotrypsin inhibition assay

Bovine chymotrypsin (Calbiochem, Darmstadt, Germany) was assayed spectrophotometrically (Varian, Cary 50 Bio) at 405 nm at 25  $^{\circ}\text{C}$  for 10 min. Inhibitor stock solutions were prepared in DMSO. Assay buffer was 20  $\text{mM}$  Tris HCl buffer, 150  $\text{mM}$  NaCl, pH 8.4. An enzyme stock solution of 10  $\mu\text{g}/\text{mL}$  was prepared in 1  $\text{mM}$  HCl and diluted with assay buffer. A 40  $\text{mM}$  stock solution of the chromogenic substrate Suc-Ala-Ala-Pro-Phe-pNA (Bachem, Bubendorf, Switzerland) in DMSO was diluted with assay buffer. The final concentration of the substrate was 200  $\mu\text{M}$ , of DMSO was 6% and of chymotrypsin was 12.5  $\text{ng}/\text{mL}$ . Into a cuvette containing 845  $\mu\text{L}$  assay buffer, 55  $\mu\text{L}$  of an inhibitor solution and 50  $\mu\text{L}$  of a substrate solution were added and thoroughly mixed. The reaction was initiated by adding 50  $\mu\text{L}$  of a chymotrypsin solution.



## Trypsin inhibition assay

Bovine pancreas trypsin (Sigma, Steinheim, Germany) was assayed spectrophotometrically (Varian, Cary 50 Bio) at 405 nm at 25 °C for 10 min. Inhibitor stock solutions were prepared in DMSO. Assay buffer was 20 mM Tris HCl buffer, 150 mM NaCl, pH 8.4. An enzyme stock solution of 10 µg/mL was prepared in 1 mM HCl and diluted with assay buffer. A 40 mM stock solution of the chromogenic substrate Suc-Ala-Ala-Pro-Arg-pNA (Bachem, Bubendorf, Switzerland) in DMSO was diluted with assay buffer. The final concentration of the substrate was 200 µM, of DMSO was 6% and of trypsin was 12.5 ng/mL. Into a cuvette containing 845 µL assay buffer, 55 µL of an inhibitor solution and 50 µL of a substrate solution were added and thoroughly mixed. The reaction was initiated by adding 50 µL of a trypsin solution.

## Acetyl cholinesterase inhibition assay

Acetylcholinesterase (AChE) from *Electrophorus electricus* (Fluka, Deisenhofen, Germany) was assayed spectrophotometrically (Varian, Cary 50 Bio) at 412 nm at 25 °C for 10 min. Inhibitor stock solutions were prepared in DMSO. Assay buffer was 100 mM sodium phosphate, 100 mM NaCl, pH 7.3. The enzyme stock solution (~100 U/mL) in assay buffer was kept at 0 °C. Appropriate dilutions were prepared immediately before starting the measurement. ATCh (Sigma, Steinheim, Germany) (10 mM) and DTNB (Sigma, Steinheim, Germany) (7 mM) were dissolved in assay buffer and kept at 0 °C. The final concentration of ATCh was 500 µM, of DTNB was 350 µM, of acetonitrile was 5%, of DMSO was 1%, and of AChE was ~30 mU/mL. Into a cuvette containing 830 µL assay buffer, 50 µL of the DTNB solution, 50 µL acetonitrile, 10 µL of the inhibitor solution, and 10 µL of an enzyme solution (~3 U/mL) were added and thoroughly mixed. After incubation for 15 min at 25 °C, the reaction was initiated by adding 50 µL of the ATCh solution.

## Cholesterol esterase inhibition assay

Cholesterol esterase (CEase) from bovine pancreas (Sigma, Steinheim, Germany) was assayed spectrophotometrically (Varian, Cary 50 Bio) at 405 nm at 25 °C for 10 min. Inhibitor stock solutions were prepared in DMSO. Assay buffer was 100 mM sodium phosphate, 100 mM NaCl, pH 7.0. An enzyme stock solution (122 µg/mL) was prepared in 100 mM sodium phosphate buffer, pH 7.0, kept at 0 °C and was diluted immediately before starting the measurement. TC (Sigma, Steinheim, Germany) (12 mM) was dissolved in assay

buffer and kept at 25 °C. A stock solution of pNPB (Sigma, Steinheim, Germany) (20 mM) was prepared in acetonitrile. The final concentration of the substrate pNPB was 200 µM, of acetonitrile was 5%, of DMSO was 1%, of TC was 6 mM, and of CEase was 10 ng/mL. Into a cuvette containing 430 µL assay buffer, 500 µL of the TC solution, 40 µL acetonitrile, 10 µL of the pNPB solution, and 10 µL of the inhibitor solution were added and thoroughly mixed. After incubation for 5 min at 25 °C, the reaction was initiated by adding 10 µL of the enzyme solution (1 µg/mL).

# Bibliography

- Abell, A. D.; Jones, M. A.; Coxon, J. M.; Morton, J. D.; Aitken, S. G.; McNabb, S. B.; Lee, H. Y.; Mehrtens, J. M.; Alexander, N. A.; Stuart, B. G.; Neffe, A. T.; Bickerstaffe, R. Molecular modeling, synthesis, and biological evaluation of macrocyclic calpain inhibitors. *Angew. Chem. Int. Ed.* **2009**, *48*, 1455–1458.
- Altmann, E.; Aichholz, R.; Betschart, C.; Buhl, T.; Green, J.; Irie, O.; Teno, N.; Lattmann, R.; Tintelnot-Blomley, M.; Missbach, M. 2-cyano-pyrimidines: a new chemotype for inhibitors of the cysteine protease cathepsin K. *J. Med. Chem.* **2007**, *50*, 591–594.
- Altmann, E.; Renaud, J.; Green, J.; Farley, D.; Cutting, B.; Jahnke, W. Arylaminoethyl amides as novel non-covalent cathepsin K inhibitors. *J. Med. Chem.* **2002**, *45*, 2352–2354.
- Bajorath, J. Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.
- Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894.
- Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating structure-activity landscapes. *Drug Discov. Today* **2009**, *14*, 698–705.
- Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.

- Batista, J.; Bajorath, J. Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 1937–1944.
- Batista, J.; Bajorath, J. Mining of randomly generated molecular fragment populations uncovers activity specific fragment hierarchies. *J. Chem. Inf. Model.* **2007**, *47*, 59–68.
- Batista, J.; Bajorath, J. Similarity searching using compound class-specific combinations of substructures found in randomly generated molecular fragment populations. *ChemMedChem* **2008**, *3*, 67–73.
- Baum, B.; Mohamed, M.; Zayed, M.; Gerlach, C.; Heine, A.; Hangauer, D.; Klebe, G. More than a simple lipophilic contact: a detailed thermodynamic analysis of nonbasic residues in the S1 pocket of thrombin. *J. Mol. Biol.* **2009**, *390*, 56–69.
- Béliveau, F.; Désilets, A.; Leduc, R. Probing the substrate specificities of matriptase, matriptase-2, hepsin and DESC1 with internally quenched fluorescent peptides. *FEBS J.* **2009**, *276*, 2213–2226.
- Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2002**, *2*, 3204–3218.
- Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004a**, *44*, 170–178.
- Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004b**, *44*, 1708–1718.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G. T.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucl. Acids Res.* **2000**, *28*, 235–242.
- Blair, H. C.; Athanasou, N. A. Recent advances in osteoclast biology and pathological bone resorption. *Histol. Histopathol.* **2004**, *19*, 189–199.
- Bode, W.; Wei, A. Z.; Huber, R.; Meyer, E.; Travis, J.; Neumann, S. X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. *EMBO J.* **1986**, *5*, 2453–2458.

- Böhm, H. J. Development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput-Aided Mol. Des.* **1994**, *8*, 243–256.
- Böhm, H. J.; Schneider, G. Introduction to molecular recognition models. In *Protein-Ligand Interactions*; Böhm, H. J., Schneider, G., Eds.; Wiley: New York, **2003**
- Böhm, H.-J.; Kleb, G. What Can We Learn from Molecular Recognition in Protein-Ligand Complexes for the Design of New Drugs? *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 2588–2614.
- Breitenlechner, C. B.; Wegge, T.; Berillon, L.; Graul, K.; Marzenell, K.; Friebe, W.G.; Thomas, U.; Schumacher, R.; Huber, R.; Engh, R. A.; Masjost, B. Structure-based optimization of novel azepane derivatives as PKB inhibitors. *J. Med. Chem.* **2004**, *47*, 1375–1390.
- Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401–3408.
- Brömme, D.; Kaleta, J. Thiol-dependent cathepsins: pathophysiological implications and recent advances in inhibitor design. *Curr. Pharm. Des.* **2002**, *8*, 1639–1658.
- Brown, R. D.; Martin, Y. C. An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR QSAR Environ Res.* **1998**, *8*, 23–39.
- Bubik, A.; Sedmak, B.; Novinec, M.; Lenarčič, B.; Lah, T. T. Cytotoxic and peptidase inhibitory activities of selected non-hepatotoxic cyclic peptides from cyanobacteria. *Biol. Chem.* **2008**, *389*, 1339–1346.
- Bugge, T. H.; Antalis, T. M.; Wu, Q. Type II transmembrane serine proteases. *J. Biol. Chem.* **2009**, *284*, 23177–23181.
- Cal, S.; Quesada, V.; Garabaya, C.; Lopez-Otin, C. Polyserase-I, a human polyprotease with the ability to generate independent serine protease domains from a single translation product. *Proc. Natl. Acad. Sci. U S A* **2003**, *100*, 9185–9190.
- Chatterjee, A. K.; Liu, H.; Tully, D. C.; Guo, J.; Epple, R.; Russo, R.; Williams, J.; Roberts, M.; Tuntland, T.; Chang, J.; Gordon, P.; Hollenbeck, T.; Tumanut, C.; Li, J.; Harris, J. L. Synthesis and SAR of succinamide peptidomimetic inhibitors of cathepsin S. *Bioorg. Med. Chem. Lett.* **2007**, *10*, 2899–2903.

- Choi, S. Y.; Bertram, S.; Glowacka, I.; Park, Y. W.; Pöhlmann, S. Type II transmembrane serine proteases in cancer and viral infections. *Trends Mol. Med.* **2009**, *15*, 303–312.
- Chua, F.; Laurent, G. J. Neutrophil elastase: mediator of extracellular matrix destruction and accumulation. *Proc. Am. Thorac. Soc.* **2006**, *3*, 424–427.
- Chuaqui, C.; Deng, Z.; Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121–133.
- Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J. Chem. Inf. Model.* **2010**, *50*, 470–479.
- Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51*, 3661–3680.
- Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead hopping". Validation of topomer similarity as a superior predictor of similar biological activities. *J. Med. Chem.* **2004**, *47*, 6777–6791.
- Crane, S. N.; Black, W. C.; Palmer, J. T.; Davis, D. E.; Setti, E.; Robichaud, J.; Paquet, J.; Oballa, R. M.; Bayly, C. I.; McKay, D. J.; Somoza, J. R.; Chauret, N.; Seto, C.; Scheigetz, J.; Wesolowski, G.; Massé, F.; Desmarais, S.; Ouellet, M. Beta-substituted cyclohexanecarboxamide: a nonpeptidic framework for the design of potent inhibitors of cathepsin K. *J. Med. Chem.* **2006**, *49*, 1066–1079.
- Crisman, T. J.; Bender, A.; Milik, M.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Fejzo, J.; Hommel, U.; Davies, J. W.; Glick, M. "Virtual fragment linking": an approach to identify potent binders from low affinity fragment hits. *J. Med. Chem.* **2008**, *51*, 2481–2491.

- Crisman, T. J.; Sisay, M. T.; Bajorath, J. Ligand-target interaction-based weighting of substructures for virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 1955–1964.
- De Domenico, I.; McVey Ward, D.; Kaplan, J. Regulation of iron acquisition and storage: consequences for iron-linked disorders. *Nat. Rev. Mol. Cell. Biol.* **2008**, *9*, 72–81.
- Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.
- Driessen, C.; Bryant, R. A. R.; Lennon-Duménil, A. -M.; Villadangos, J. A.; Bryant, P. W.; Shi, G. -P.; Chapman, H. A.; Ploegh, H. L. Cathepsin S controls the trafficking and maturation of MHC class II molecules in dendritic cells. *J. Cell Biol.* **1999**, *147*, 775–790.
- Du, X.; She, E.; Gelbart, T.; Truksa, J.; Lee, P.; Xia, Y.; Khovananth, K.; Mudd, S.; Mann, N.; Moresco, E. M. Y.; Beutler, E.; Beutler, B. The serine protease TMPRSS6 is required to sense iron deficiency. *Science* **2008**, *320*, 1088–1092.
- Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
- Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **2007**, *12*, 225–233.
- Eckert, H.; Vogt, I.; Bajorath, J. Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: design of DynaMAD and comparison with MAD and DMC. *J. Chem. Inf. Model.* **2006**, *46*, 1623–1634.
- Eilfeld, A.; González Tanarro, C. M.; Frizler, M.; Sieler, J.; Schulze, B.; Gütschow, M. Synthesis and elastase-inhibiting activity of 2-pyridinyl-isothiazol-3(2H)-one 1,1-dioxides. *Bioorg. Med. Chem.* **2008**, *16*, 8127–8135.
- Falgueyret, J. P.; Desmarais, S.; Oballa, R.; Black, W. C.; Cromlish, W.; Khougaz, K.; Lamontagne, S.; Massé, F.; Riendeau, D.; Toulmond, S.; Percival, M. D. Lysosomotropism of basic cathepsin K inhibitors contributes to increased cellular potencies against off-target cathepsins and reduced functional selectivity. *J. Med. Chem.* **2005**, *48*, 7535–7543.

- Finberg, K. E.; Heeney, M. M.; Campagna, D. R.; Ayşinok, Y.; Pearson, H. A.; Hartman, K. R.; Mayo, M. M.; Samuel, S. M.; Strouse, J. J.; Markianos, K.; Andrews, N. C.; Fleming, M. D. Mutations in TMPRSS6 cause iron-refractory iron deficiency anaemia (IRIDA). *Nat. Genet.* **2008**, *40*, 569–571.
- Folgueras, A. R.; de Lara, F. M.; Pendás, A. M.; Garabaya, C.; Rodriguez, F.; Astudillo, A.; Bernal, T.; Cabanillas, R.; López-Otín, C.; Velasco, G. Membrane-bound serine protease matriptase-2 (Tmprss6) is an essential regulator of iron homeostasis. *Blood* **2008**, *112*, 2539–2545.
- Fox, S.; Farr-Jones, S.; Sopchak, L.; Boggs, A.; Nicely, H. W.; Khoury, R.; Biro, M. High-throughput screening: update on practices and success. *J. Biomol. Screen.* **2006**, *1*, 864–869.
- Friedrich, R.; Fuentes-Prior, P.; Ong, E.; Coombs, G.; Hunter, M.; Oehler, R.; Pierson, D.; Gonzalez, R.; Huber, R.; Bode, W.; Madison, E. L. Catalytic domain structures of MT-SP1/matriptase, a matrix-degrading transmembrane serine proteinase. *J. Biol. Chem.* **2002**, *277*, 2160–2168.
- Frizler, M.; Stirnberg, M.; Sisay, M. T.; Gütschow, M. Development of nitrile-based peptidic inhibitors of cysteine cathepsins. *Curr. Top. Med. Chem.* **2010**, *10*, 294–322.
- Fujii, K.; Sivonen, K.; Adachi, K.; Noguchi, K.; Sano, H.; Hirayama, K.; Suzuki, M.; Harada, K. I. Comparative study of toxic and non-toxic cyanobacterial products: Novel peptides from toxic *Nodularia spumigena* AV1. *Tetrahedron Lett.* **1997**, *38*, 5525–5528.
- Gauthier, J. Y.; Black, W. C.; Courchesne, I.; Cromlish, W.; Desmarais, S.; Houle, R.; Lamontagne, S.; Li, C. S.; Massé, F.; McKay, D. J.; Ouellet, M.; Robichaud, J.; Truchon, J.-F.; Truong, V.-L.; Wang, Q.; Percival, M. D. The identification of potent, selective, and bioavailable cathepsin S inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4929–4933.
- Gauthier, J. Y.; Chauret, N.; Cromlish, W.; Desmarais, S.; Duong le, T.; Falgout, J. P.; Kimmel, D. B.; Lamontagne, S.; Léger, S.; LeRiche, T.; Li, C. S.; Massé, F.; McKay, D. J.; Nicoll-Griffith, D. A.; Oballa, R. M.; Palmer, J. T.; Percival, M. D.; Riendeau, D.; Robichaud, J.; Rodan, G. A.; Rodan, S. B.; Seto, C.; Thérien, M.; Truong, V. L.; Venuti, M. C.; Wesolowski, G.; Young, R. N.; Zamboni, R.; Black, W. C. The discovery of odanacatib (MK-0822), a selective inhibitor of cathepsin K. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 923–928.



- Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.
- Gershell, L. J.; & Atkins, J. H. A brief history of novel drug discovery technologies. *Nat. Rev. Drug Discov.* **2003**, *2*, 321–327.
- Ghosh, S.; Nie, A.; An, J.; Huang, Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.* **2006**, *10*, 194–202.
- Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165–179.
- Greco, M. N.; Powell, E. T.; Hecker, L. R.; Andrade-Gordon, P.; Kauffman, J. A.; Lewis, J. M.; Ganesh, V.; Tulinsky, A.; Maryanoff, B. E. Novel thrombin inhibitors that are based on a macrocyclic tripeptide motif. *Bioorg. Med. Chem. Lett.* **1996**, *6*, 2947–2952.
- Griffith, R.; Luu, T. T.; Garner, J.; Keller, P. Combining structure-based drug design and pharmacophores. *J. Mol. Graph. Model.* **2005**, *23*, 439–446.
- Guha, R.; Van Drie, J. H. Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.
- Guillem, F.; Lawson, S.; Kannengiesser, C.; Westerman, M.; Beaumont, C.; Grandchamp, B. Two nonsense mutations in the TMPRSS6 gene in a patient with microcytic anemia and iron deficiency. *Blood* **2008**, *112*, 2089–2091.
- Gustin, D. J.; Schon, C. A.; Wei, J.; Cai, H.; Meduna, S. P.; Khatuya, H.; Sun, S.; Gu, Y.; Jiang, W.; Thurmond, R. L.; Karlsson, L.; Edwards, J. P. Discovery and SAR studies of a novel series of noncovalent cathepsin S inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1687–1691.
- Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons. *Nature Rev. Drug Discov.* **2007**, *6*, 211–219.
- Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.

- Hellstern, P.; Stürzebecher, U.; Wuchold, B.; Haubelt, H.; Seyfert, U. T.; Bauer, M.; Vogt, A.; Stürzebecher, J. Preservation of in vitro function of platelets stored in the presence of a synthetic dual inhibitor of factor Xa and thrombin. *J. Thromb. Haemost.* **2007**, *5*, 2119–2126.
- Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004a**, *2*, 3256–3266.
- Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004b**, *44*, 1177–1185.
- Honey, K.; Rudensky, A. Y. Lysosomal cysteine proteases regulate antigen presentation. *Nat. Rev. Immunol.* **2003**, *3*, 472–482.
- Honkanen, R. E.; Zwiller, J.; Moore, R. E.; Daily, S. L.; Khatra, B. S.; Dukelow, M.; Boynton, A. L. Characterization of microcystin-LR, a potent inhibitor of type 1 and type 2A protein phosphatases. *J. Biol. Chem.* **1990**, *265*, 19401–19404.
- Hooper, J. D.; Clements, J. A.; Quigley, J. P.; Antalis, T. M. Type II transmembrane serine proteases. Insights into an emerging class of cell surface proteolytic enzymes. *J. Biol. Chem.* **2001**, *276*, 857–860.
- Inagaki, H., Tsuruoka, H., Hornsby, M., Lesley, S. A., Spraggon, G., and Ellman, J. A. Characterization and optimization of selective, non-peptidic inhibitors of cathepsin S with an unprecedented binding mode. *J. Med. Chem.* **2007**, *50*, 2693–2699.
- Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
- Itou, Y.; Suzuki, S.; Ishida, K.; Murakami, M. Anabaenopeptins G and H, potent carboxypeptidase A inhibitors from the cyanobacterium *Oscillatoria agardhii* (NIES-595). *Bioorg. Med. Chem. Lett.* **1999**, *9*, 1243–1246.
- Jorgensen, A.; Langgard, M.; Gundertofte, K.; Pedersen, J. T. A fragment-weighted key-based similarity measure for use in structural clustering and virtual screening. *QSAR Comb. Sci.* **2006**, *3*, 221–234.

- Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–1818.
- Katunuma, N.; Matsunaga, Y.; Himeno, K.; Hayashi, Y. Insights into the roles of cathepsins in antigen processing and presentation revealed by specific inhibitors. *Biol. Chem.* **2003**, *384*, 883–890.
- Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942–1951.
- Kim, R.; Skolnick, J. Assessment of programs for ligand binding affinity prediction. *J. Comput. Chem.* **2008**, *29*, 1316–1331.
- Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.
- Kodama, T.; Yukioka, H.; Kato, T.; Kato, N.; Hato, F.; Kitagawa, S. Neutrophil elastase as a predicting factor for development of acute lung injury. *Intern. Med.* **2007**, *46*, 699–704.
- Kodani, S.; Suzuki, S.; Ishida, K.; Murakami, M. Five new cyanobacterial peptides from water bloom materials of lake Teganuma (Japan). *FEMS Microbiol. Lett.* **1999**, *178*, 343–348.
- Korkmaz, B.; Moreau, T.; Gauthier, F. Neutrophil elastase, proteinase 3 and cathepsin G: physicochemical properties, activity and physiopathological functions. *Biochimie* **2008**, *90*, 227–242.
- Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins Struct. Funct. Genet.* **1999**, *37*, 228–241.
- Kubinyi, H. The changing landscape in drug discovery, in: computational approaches to structure based drug design, R. M. Stroud, Ed., Royal Society of Chemistry, London, **2007**, 24–45.
- Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule - ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
- Langer, T.; Hoffmann, R.; Bryant, S.; Lesur, B. Hit finding: towards 'smarter' approaches. *Curr. Opin. Pharmacol.* **2009**, *9*, 589–593.

- Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.
- Lee, A. Y.; Smitka, T. A.; Bonjouklian, R.; Clardy, J. Atomic structure of the trypsin-A90720A complex: a unified approach to structure and function. *Chem. Biol.* **1994**, *1*, 113–117.
- Lee, M. S. Matrix-degrading type II transmembrane serine protease matriptase: its role in cancer development and malignancy. *J. Cancer Mol.* **2006**, *2*, 183–190.
- Lee, S. L.; Dickson, R. B.; Lin, C. Y. Activation of hepatocyte growth factor and urokinase plasminogen activator by matriptase, an epithelial membrane serine protease. *J. Biol. Chem.* **2000**, *275*, 36720–36725.
- Leeson, P. D.; Davis, A. M.; Steele, J. Drug-like properties: guiding principles for design – or chemical prejudice? *Drug. Discov. Today. Technol.* **2004**, *1*, 189–195.
- Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discov. Today* **2004**, *9*, 27–34.
- Leung, D.; Abbenante, G.; Fairlie, D. P. Protease inhibitors: current status and future prospects. *J. Med. Chem.* **2000**, *43*, 305–341.
- Leung-Toung, R.; Zhao, Y.; Li, W.; Tam, T. F.; Karimian, K.; Spino, M. Thiol proteases: inhibitors and potential therapeutic targets. *Curr. Med. Chem.* **2006**, *13*, 547–581.
- Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
- Lin, C. Y.; Anders, J.; Johnson, M.; Sang, Q. A.; Dickson, R. B. Molecular cloning of cDNA for matriptase, a matrix-degrading serine protease with trypsin-like activity. *J. Biol. Chem.* **1999**, *274*, 18231–18236.
- Lin, J.; Sahakian, D. C.; de Morais, S. M.; Xu, J. J.; Polzer, R. J. et al. The role of absorption, distribution, metabolism, excretion and toxicity in drug discovery. *Curr. Top. Med. Chem.* **2003**, *3*, 1125–1154.

- Linington, R. G.; Edwards, D. J.; Shuman, C. F.; McPhail, K. L.; Matainaho, T.; Gerwick, W. H. Symplocamide A, a potent cytotoxin and chymotrypsin inhibitor from the marine Cyanobacterium *Symploca* spp. *J. Nat. Prod.* **2008**, *71*, 22–27.
- Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug deliv. Rev.* **2001**, *46*, 3–26.
- List, K.; Szabo, R.; Wertz, P. W.; Segre, J.; Haudenschild, C. C.; Kim, S. Y.; Bugge, T. H. Loss of proteolytically processed filaggrin caused by epidermal deletion of Matriptase/MT-SP1. *J. Cell Biol.* **2003**, *163*, 901–910.
- Liu, H.; Tully, D. C.; Epple, R.; Bursulaya, B.; Li, J.; Harris, J. L.; Williams, J. A.; Russo, R.; Tumanut, C.; Roberts, M. J.; Alper, P. B.; He, Y.; Karanewsky, D. S. Design and synthesis of arylaminoethyl amides as noncovalent inhibitors of cathepsin S. Part 1. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 4979–4984.
- Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucl. Acids. Res.* **2007**, *35*, D198–D201.
- Löser, R.; Abbenante, G.; Madala, P. K.; Halili, M.; Le, G. T.; Fairlie, D. P. Noncovalent tripeptidyl benzyl- and cyclohexyl-amine inhibitors of the cysteine protease caspase-1. *J. Med. Chem.* **2010**, *53*, 2651–2655.
- Löser, R.; Frizler, M.; Schilling, K.; Gütschow, M. Azadipeptide nitriles: highly potent and proteolytically stable inhibitors of papain-like cysteine proteases. *Angew. Chem. Int. Ed.* **2008**, *47*, 4331–4334.
- Löser, R.; Schilling, K.; Dimmig, E.; Gütschow, M. Interaction of papain-like cysteine proteases with dipeptide-derived nitriles. *J. Med. Chem.* **2005**, *48*, 7688–7707.
- Loughlin, W. A.; Tyndall, J. D.; Glenn, M. P.; Fairlie, D. P. Beta-strand mimetics. *Chem. Rev.* **2004**, *104*, 6085–6117.
- Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today*, **2002**, *7*, 1047–1055.

- MacFaul, P. A.; Morley, A. D.; Crawford, J. J. A simple in vitro assay for assessing the reactivity of nitrile containing compounds. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 1136–1138.
- MacKintosh, C.; Beattie, K. A.; Klumpp, S.; Cohen, P.; Codd, G. A. Cyanobacterial microcystin-LR is a potent and specific inhibitor of protein phosphatases 1 and 2A from both mammals and higher plants. *FEBS Lett.* **1990**, *264*, 187–192.
- Maggiora, G. M. On outliers and activity cliffs - why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
- Malyszko, J. Hemojuvelin: the hepcidin story continues. *Kidney Blood Press. Res.* **2009**, *32*, 71–76.
- Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195–207.
- Markt, P.; McGoochan, C.; Walker, B.; Kirchmair, J.; Feldmann, C.; De Martino, G.; Spitzer, G.; Distinto, S.; Schuster, D.; Wolber, G.; Laggner, C.; Langer, T. Discovery of novel cathepsin S inhibitors by pharmacophore-based virtual high-throughput screening. *J. Chem. Inf. Model.* **2008**, *48*, 1693–1705.
- Marquis, R. W.; Ru, Y.; LoCastro, S. M.; Zeng, J.; Yamashita, D. S.; Oh, H. J.; Erhard, K. F.; Davis, L. D.; Tomaszek, T. A.; Tew, D.; Salyers, K.; Proksch, J.; Ward, K.; Smith, B.; Levy, M.; Cummings, M. D.; Haltiwanger, R. C.; Trescher, G.; Wang, B.; Hemling, M. E.; Quinn, C. J.; Cheng, H. Y.; Lin, F.; Smith, W. W.; Janson, C. A.; Zhao, B.; McQueney, M. S.; D'Alessio, K.; Lee, C. P.; Marzulli, A.; Dodds, R. A.; Blake, S.; Hwang, S. M.; James, I. E.; Gress, C. J.; Bradley, B. R.; Lark, M. W.; Gowen, M.; Veber, D. F. Azepanone-based inhibitors of human and rat cathepsin K. *J. Med. Chem.* **2001**, *44*, 1380–1395.
- Marsh, I. R.; Bradley, M.; Teague, S. J. Solid-phase total synthesis of Oscillamide Y and analogues. *J. Org. Chem.* **1997**, *62*, 6199–6203.
- Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350–4358.

- Matern, U.; Schleberger, C.; Jelakovic, S.; Weckesser, J.; Schulz, G. E. Binding structure of elastase inhibitor scyptolin A. *Chem. Biol.* **2003**, *10*, 997–1001.
- Mausser, H.; Stahl, M. Chemical fragment spaces for de novo design. *J. Chem. Inf. Model.* **2007**, *47*, 318–324.
- Mayr, L. M.; Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580–588.
- McDonough, M. A.; Schofield, C. J. New structural insights into the inhibition of serine proteases by cyclic peptides from bacteria. *Chem. Biol.* **2003**, *10*, 898–900.
- McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
- Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477–491.
- Mehner, C.; Müller, D.; Kehraus, S.; Hautmann, S.; Gütschow, M.; König, G. M. New peptolides from the cyanobacterium *Nostoc insulare* as selective and potent inhibitors of human leukocyte elastase. *ChemBioChem* **2008**, *9*, 2692–2703.
- Mehrotra, A. P.; Webster, K. L.; Gani, D. Design and preparation of serine-threonine protein phosphatase inhibitors based upon the nodularin and microcystin toxin structures: Part 1. Evaluation of key inhibitory features and synthesis of a rationally stripped-down nodularin macrocycle. *J. Chem. Soc. Perkin Trans. 1* **1997**, 2495–2511.
- Melis, M. A.; Cau, M.; Congiu, R.; Sole, G.; Barella, S.; Cao, A.; Westerman, M.; Cazzola, M.; Galanello, R. A mutation in the *TMPRSS6* gene, encoding a transmembrane serine protease that suppresses hepcidin production, in familial iron deficiency anemia refractory to oral iron. *Haematologica* **2008**, *93*, 1473–1479.
- Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comp. Chem.* **1992**, *13*, 505–524.

- Merlot, C.; Domine, D.; Cleve, C.; Church, D. J. Chemical substructures in drug discovery. *Drug Discov. Today* **2003**, *8*, 594–602.
- Mestres, J. Virtual screening: a real screening complement to high-throughput screening. *Biochem. Soc. Trans.* **2002**, *30*, 797–799.
- Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153*, S7–S26.
- Moore, R. E. Cyclic peptides and depsipeptides from cyanobacteria: a review. *J. Ind. Microbiol.* **1996**, *16*, 134–143.
- Müller, D.; Krick, A.; Kehraus, S.; Mehner, C.; Hart, M.; Küpper, F. C.; Saxena, K.; Prinz, H.; Schwalbe, H.; Janning, P.; Waldmann, H.; König, G. M. Brunsvicamides A-C: sponge-related cyanobacterial peptides with Mycobacterium tuberculosis protein tyrosine phosphatase inhibitory activity. *J. Med. Chem.* **2006**, *49*, 4871–4878.
- Murakami, M.; Suzuki, S.; Itou, Y.; Kodani, S.; Ishida, K. New anabaenopeptins, potent carboxypeptidase-A inhibitors from the cyanobacterium *Aphanizomenon flos-aquae*. *J. Nat. Prod.* **2000**, *63*, 1280–1282.
- Nägler, D. K.; Ménard, R. Family C1 cysteine proteases: biological diversity or redundancy? *Biol. Chem.* **2003**, *384*, 837–843.
- Nelson, T. D.; LeBlond, C. R.; Frantz, D. E.; Matty, L.; Mitten, J. V.; Weaver, D. G.; Moore, J. C.; Kim, J. M.; Boyd, R.; Kim, P. -Y.; Gbewonyo, K.; Brower, M.; Sturr, M.; McLaughlin, K.; McMasters, D. R.; Kress, M. H.; McNamara, J. M.; Dolling, U. H. Stereoselective synthesis of a potent thrombin inhibitor by a novel P2-P3 lactone ring opening. *J. Org. Chem.* **2004**, *69*, 3620–3627.
- Nemeth, E.; Tuttle, M. S.; Powelson, J.; Vaughn, M. B.; Donovan, A.; Ward, D. M.; Ganz, T.; Kaplan, J. Heparin regulates cellular iron efflux by binding to ferroportin and inducing its internalization. *Science* **2004**, *306*, 2090–2093.
- Netzel-Arnett, S.; Hooper, J. D.; Szabo, R.; Madison, E. L.; Quigley, J. P.; Bugge, T. H.; Antalis, T. M. Membrane anchored serine proteases: A rapidly expanding group of cell surface proteolytic enzymes with potential roles in cancer. *Cancer Metastasis Rev.* **2003**, *22*, 237–258.



- Niederkofler, V.; Salie, R.; Arber, S. Hemojuvelin is essential for dietary iron sensing, and its mutation leads to severe iron overload. *J. Clin. Invest.* **2005**, *115*, 2180–2186.
- Noel, A.; Maillard, C.; Rocks, N.; Jost, M.; Chabottaux, V.; Sounni, N. E.; Maquoi, E.; Cataldo, D.; Foidart, J. M. Membrane associated proteases and their inhibitors in tumour angiogenesis. *J. Clin. Pathol.* **2004**, *57*, 577–584.
- Oballa, R. M.; Truchon, J. -F.; Bayly, C. I.; Chauret, N.; Day, S.; Crane, S.; Berthelette, C. A generally applicable method for assessing the electrophilicity and reactivity of diverse nitrile-containing compounds. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 998–1002.
- Oberst, M. D.; Johnson, M. D.; Dickson, R. B.; Lin, C. Y.; Singh, B.; Stewart, M.; Williams, A.; Nafussi, A.; Smyth, J. F.; Gabra, H.; Sellar, G. C. Expression of the serine protease matriptase and its inhibitor HAI-1 in epithelial ovarian cancer: correlation with clinical outcome and tumor clinicopathological parameters. *Clin. Cancer. Res.* **2002**, *8*, 1101–1107.
- Okumura, Y.; Hayama, M.; Takahashi, E.; Fujiuchi, M.; Shimabukuro, A.; Yano, M.; Kido, H. Serase-1B, a new splice variant of polyserase-1/TMPRSS9, activates urokinase-type plasminogen activator and the proteolytic activation is negatively regulated by glycosaminoglycans. *Biochem J.* **2006**, *400*, 551–561.
- Oprea, T. I. Virtual screening in lead discovery: a viewpoint. *Molecules* **2002**, *7*, 51–62.
- Palmer, J. T.; Bryant, C.; Wang, D.-X.; Davis, D. E.; Setti, E. L.; Rydzewski, R. M.; Venkatraman, S.; Tian, Z.-Q.; Burrill, L. C.; Mendonca, R. V.; Springman, E.; McCarter, J.; Chung, T.; Cheung, H.; Janc, J. W.; McGrath, M.; Somoza, J. R.; Enriquez, P.; Yu, Z. W.; Strickley, R. M.; Liu, L.; Venuti, M. C.; Percival, M. D.; Falgoutyret, J.-P.; Prasit, P.; Oballa, R.; Riendeau, D.; Young, R. N.; Wesolowski, G.; Rodan, S. B.; Johnson, C.; Kimmel, D. B.; Rodan, G. Design and synthesis of tri-ring P3 benzamide-containing aminonitriles as potent, selective, orally effective inhibitors of cathepsin K. *J. Med. Chem.* **2005**, *48*, 7520–7534.

- Papanikolaou, G.; Samuels, M. E.; Ludwig, E. H.; MacDonald, M. L.; Franchini, P. L.; Dubé, M. P.; Andres, L.; MacFarlane, J.; Sakellaropoulos, N.; Politou, M.; Nemeth, E.; Thompson, J.; Risler, J. K.; Zaborowska, C.; Babakaiff, R.; Radomski, C. C.; Pape, T. D.; Davidas, O.; Christakis, J.; Brissot, P.; Lockitch, G.; Ganz, T.; Hayden, M. R.; Goldberg, Y. P. Mutations in HFE2 cause iron overload in chromosome 1q-linked juvenile hemochromatosis. *Nat. Genet.* **2004**, *36*, 77–82.
- Park, T. J.; Lee, Y. J.; Kim, H. J.; Park, H. G.; Park, W. J. Cloning and characterization of TMPRSS6, a novel type 2 transmembrane serine protease. *Mol. Cells* **2005**, *19*, 223–227.
- Parr, C.; Sanders, A. J.; Davies, G.; Martin, T.; Lane, J.; Mason, M. D.; Mansel, R. E.; Jiang, W. G. Matriptase-2 inhibits breast growth and invasion and correlates with favorable prognosis for breast cancer patients. *Clin. Cancer. Res.* **2007**, *13*, 3568–3576.
- Patterson, A. W.; Wood, W. J. L.; Hornsby, M.; Lesley, S.; Spraggon, G.; Ellman, J. A. Identification of selective, nonpeptidic nitrile inhibitors of cathepsin S using the substrate activity screening method. *J. Med. Chem.* **2006**, *49*, 6298–6307.
- Peltason L.; Bajorath J. SAR Index: quantifying the nature of a structure-activity Relationships. *J. Med. Chem.* **2007a**, *50*, 5571–5578.
- Peltason, L.; Bajorath, J. Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chem. Biol.* **2007b**, *14*, 489–497.
- Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of structure-activity relationship determinants in analogue Series. *J. Med. Chem.* **2009**, *52*, 3212–3224.
- Peltason, L.; Bajorath, J. Systematic computational analysis of structure-activity relationships: concepts, challenges and recent advances. *Future Med. Chem.* **2009**, *1*, 451–466.
- Pérez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixidó, J. APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 1245–1260.

- Pham, C. T. Neutrophil serine proteases: specific regulators of inflammation. *Nat. Rev. Immunol.* **2006**, *6*, 541–550.
- Pietsch, M.; Gütschow, M. Alternate substrate inhibition of cholesterol esterase by thieno[2,3-d][1,3]oxazin-4-ones. *J. Biol. Chem.* **2002**, *277*, 24006–24013.
- Potashman, M. H.; Duggan, M. E. Covalent modifiers: An orthogonal approach to drug design. *J. Med. Chem.* **2009**, *52*, 1231–1246.
- Ramjee, M. K.; Flinn, N. S.; Pemberton, T. P.; Quibell, M.; Wang, Y.; Watts, J. P. Substrate mapping and inhibitor profiling of falcipain-2, falcipain-3 and berghepain-2: implications for peptidase anti-malarial drug discovery. *Biochem. J.* **2006**, *399*, 47–57.
- Ramsay, A. J.; Hooper, J. D.; Folgueras, A. R.; Velasco, G.; Lopez-Otin, C. Matriptase-2 (TMPRSS6): a proteolytic regulator of iron homeostasis. *Haematologica* **2009**, *94*, 840–849.
- Ramsay, A. J.; Reid, J. C.; Velasco, G.; Quigley, J. P.; Hooper, J. D. The type II transmembrane serine protease matriptase-2 - identification, structural features, enzymology, expression pattern and potential roles. *Front. Biosci.* **2008**, *13*, 569–579.
- Rarey, M.; Kramer, B.; Lengauer, T., J. Multiple automatic base selection: protein-ligand docking based on incremental construction without manual intervention. *Comput. Aided Mol. Design.* **1997**, *11*, 369–384.
- Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–489.
- Ravikumar, M.; Pavan, S.; Bairy, S.; Pramod, A.; Sumakanth, M.; Kishore, M.; Sumithra, T. Virtual screening of cathepsin K inhibitors using docking and pharmacophore models. *Chem. Biol. Drug. Des.* **2008**, *72*, 79–90.
- Riese, R. J.; Wolf, P. R.; Brömme, D.; Natkin, L. R.; Villadangos, J. A.; Ploegh, H. L.; Chapman, H. A. Essential role for cathepsin S in MHC class II-associated invariant chain processing and peptide loading. *Immunity* **1996**, *4*, 357–366.

- Rittle, K. E.; Barrow, J. C.; Cutrona, K. J.; Glass, K. L.; Krueger, J. A.; Kuo, L. C.; Lewis, S. D.; Lucas, B. J.; McMasters, D. R.; Morrisette, M. M.; Nantermet, P. G.; Newton, C. L.; Sanders, W. M.; Yan, Y.; Vacca, J. P.; Selnick, H. G. Unexpected enhancement of thrombin inhibitor potency with *o*-aminoalkylbenzylamides in the P1 position. *Bioorg. Med. Chem. Lett.* **2003**, *20*, 3477–3482.
- Sali, A.; Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.
- Sanders, A. J.; Parr, C.; Martin, T. A.; Lane, J.; Mason, M. D.; Jiang, W. G. Genetic upregulation of matriptase-2 reduces the aggressiveness of prostate cancer cells in vitro and in vivo and affects FAK and paxillin localisation. *J. Cell. Physiol.* **2008**, *216*, 780–789.
- Sano, T.; Usui, T.; Ueda, K.; Osada, H.; Kaya, K. Isolation of new protein phosphatase inhibitors from two cyanobacteria species, *Planktothrix* spp. *J. Nat. Prod.* **2001**, *64*, 1052–1055.
- Sarabia, F.; Chammaa, S.; Ruiz, A. S.; Ortiz, L. M.; Herrera, F. J. Chemistry and biology of cyclic depsipeptides of medicinal and biological interest. *Curr. Med. Chem.* **2004**, *11*, 1309–1332.
- Schmidt, E. W.; Harper, M. K.; Faulkner, D. J. Mozamides A and B, cyclic peptides from a Theonellid sponge from Mozambique. *J. Nat. Prod.* **1997**, *60*, 779–782.
- Schneider, G.; Böhm, H. J. Virtual screening and fast automated docking methods. *Drug Discov. Today* **2002**, *7*, 64–70.
- Schneider, G.; Schneider, P. Navigation in chemical space: Ligand-based design of focused compound libraries. In: Kubinyi, H.; Müller, H. (Eds), *Chemogenomics in Drug Discovery*. Wiley-VCH, Weinheim, **2004**, 341–376.
- Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
- Schulz-Gasch, T.; Stahl, M. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discov. Today* **2004**, *1*, 231–239.

- Schweinitz, A.; Steinmetzer, T.; Banke, I. J.; Arlt, M. J. E.; Stürzebecher, A.; Schuster, O.; Geissler, A.; Giersiefen, H.; Zeslawska, E.; Jacob, U.; Krüger, A.; Stürzebecher, J. Design of novel and selective inhibitors of urokinase-type plasminogen activator with improved pharmacokinetic properties for use as antimetastatic agents. *J. Biol. Chem.* **2004**, *279*, 33613–33622.
- Schweinitz, A.; Stürzebecher, A.; Stürzebecher, U.; Schuster, O.; Stürzebecher, J.; Steinmetzer, T. New substrate analogue inhibitors of factor Xa containing 4-amidinobenzylamide as P1 residue: Part 1. *Med. Chem.* **2006**, *2*, 349–361.
- Seifert, M. H. Targeted scoring functions for virtual screening. *Drug Discov. Today*, **2009**, *14*, 562–569.
- Shi, Y. E.; Torri, J.; Yieh, L.; Wellstein, A.; Lippman, M. E.; Dickson, R. B. Identification and characterization of a novel matrix-degrading protease from hormone-dependent human breast cancer cells. *Cancer Res.* **1993**, *53*, 1409–1415.
- Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.
- Siedle, B.; Hrenn, A.; Merfort, I. Natural compounds as inhibitors of human neutrophil elastase. *Planta Med.* **2007**, *73*, 401–420.
- Silvestri, L.; Pagani, A.; Nai, A.; De Domenico, I.; Kaplan, J.; Camaschella, C. The serine protease matriptase-2 (TMPRSS6) inhibits hepcidin activation by cleaving membrane hemojuvelin. *Cell Metab.* **2008**, *8*, 502–511.
- Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural interaction fingerprints: a new approach to organizing, mining, analyzing, and designing protein-small molecule complexes. *Chem. Biol. Drug Des.* **2006**, *67*, 5–12.
- Sisay, M. T. Homology modeling and structural analysis of the catalytic domain of matriptase-2. Diploma Thesis, University of Bonn, **2007**.
- Sisay, M. T.; Hautmann, S.; Mehner, C.; König, G. M.; Bajorath, J.; Gütschow, M. Inhibition of human leukocyte elastase by brunsvicamides A-C: cyanobacterial cyclic peptides. *ChemMedChem* **2009b**, *4*, 1425–1429.

- Sisay, M. T.; Peltason, L.; Bajorath, J. Structural interpretation of activity cliffs revealed by systematic analysis of structure-activity relationships in analog series. *J. Chem. Inf. Model.* **2009a**, *49*, 2179–2189.
- Sisay, M. T.; Steinmetzer, T.; Stirnberg, M.; Mauer, E.; Hammami, M.; Bajorath, J.; Gütschow, M. Identification of the first low molecular weight inhibitors of matriptase-2. *J. Med. Chem.* **2010** (*in revision*).
- Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins* **2006**, *65*, 15–26.
- Sperandio, O.; Miteva, M. A.; Delfaud, F.; Villoutreix, B. O. Receptor-based computational screening of compound databases: the main docking-scoring engines. *Curr. Protein. Pept. Sci.* **2006**, *7*, 369–393.
- Stahl, M.; Rarey, M.; Klebe, G. Screening of databases. *Bioinformatics - From Genomes to Drugs*; VCH-Wiley Verlag GmbH: Weinheim, **2002**; 137–170.
- Stauffer, K. J.; Williams, P. D.; Selnick, H. G.; Nantermet, P. G.; Newton, C. L.; Homnick, C. F.; Zrada, M. M.; Lewis, S. D.; Lucas, B. J.; Krueger, J. A.; Pietrak, B. L.; Lyle, E. A.; Singh, R.; Miller-Stein, C.; White, R. B.; Wong, B.; Wallace, A. A.; Sitko, G. R.; Cook, J. J.; Holahan, M. A.; Stranieri-Michener, M.; Leonard, Y. M.; Lynch, J. J. Jr., McMasters, D. R.; Yan, Y. 9-Hydroxyazafluorenes and their use in thrombin inhibitors. *J. Med. Chem.* **2005**, *48*, 2282–2293.
- Steinmetzer, T.; Dönnecke, D.; Korsonewski, M.; Neuwirth, C.; Steinmetzer, P.; Schulze, A.; Saupe, S. M.; Schweinitz, A. Modification of the N-terminal sulfonyl residue in 3-amidinophenylalanine-based matriptase inhibitors. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 67–73.
- Steinmetzer, T.; Schweinitz, A.; Stürzebecher, A.; Dönnecke, D.; Uhland, K.; Schuster, O.; Steinmetzer, P.; Müller, F.; Friedrich, R.; Than, M. E.; Bode, W.; Stürzebecher, J. Secondary amides of sulfonylated 3-amidinophenylalanine. New potent and selective inhibitors of matriptase. *J. Med. Chem.* **2006**, *49*, 4116–4126.
- Stoch, S. A.; Wagner, J. A. Cathepsin K inhibitors: a novel target for osteoporosis therapy. *Clin. Pharmacol. Ther.* **2008**, *83*, 172–176.
- Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846–854.

- Stumpfe, D. Methods for computer-aided chemical biology: Exploration of compound selectivity. Doctoral thesis, University of Bonn, **2009**.
- Stumpfe, D.; Ahmed, H. E. A.; Vogt, I.; Bajorath, J. Methods for computer-aided chemical biology. Part 1: Design of a benchmark system for the evaluation of compound selectivity. *Chem. Biol. Drug. Des.* **2007**, *70*, 182–194.
- Stumpfe, D.; Frizler, M.; Sisay, M. T.; Batista, J.; Vogt, I.; Gütschow, M.; Bajorath, J. Hit expansion through computational selectivity searching. *ChemMedChem* **2009**, *4*, 52–54.
- Stumpfe, D.; Geppert, H.; Bajorath, J. Methods for computer-aided chemical biology. Part 3: analysis of structure-selectivity relationships through single- or dual-step selectivity searching and Bayesian classification. *Chem. Biol. Drug. Des.* **2008**, *71*, 518–528.
- Stumpfe, D.; Sisay, M. T.; Frizler, M.; Vogt, I.; Gütschow, M.; Bajorath, J. Inhibitors of cathepsins K and S identified using the DynaMAD virtual screening algorithm. *ChemMedChem* **2010**, *5*, 61–64.
- Stürzebecher, A.; Dönnecke, D.; Schweinitz, A.; Schuster, O.; Steinmetzer, P.; Stürzebecher, U.; Kotthaus, J.; Clement, B.; Stürzebecher, J.; Steinmetzer, T. Highly potent and selective substrate analogue factor Xa inhibitors containing D-homophenylalanine analogues as P3 residue: part 2. *ChemMedChem* **2007**, *2*, 1043–1053.
- Szabo, R.; Bugge, T. H. Type II transmembrane serine proteases in development and disease. *Int. J. Biochem. Cell Biol.* **2008**, *40*, 1297–1316.
- Szabo, R.; Netzel-Arnett, S.; Hobson, J. P.; Antalis, T. M.; Bugge, T. H. Matriptase-3 is a novel phylogenetically preserved membrane-anchored serine protease with broad serpin reactivity. *Biochem J.* **2005**, *390*, 231–242.
- Szabo, R.; Wu, Q.; Dickson, R. B.; Netzel-Arnett, S.; Antalis, T. M.; Bugge, T. H. Type II transmembrane serine proteases. *Thromb. Haemost.* **2003**, *90*, 185–193.
- Taft, C. A.; Da Silva, V. B.; Da Silva, C. H. Current topics in computer-aided drug design. *J. Pharm. Sci.* **2008**, *97*, 1089–1098.

- Taggart, C. C.; Greene, C. M.; Carroll, T. P.; O'Neill, S. J.; McElvaney, N. G. Elastolytic proteases: inflammation resolution and dysregulation in chronic infective lung disease. *Am. J. Respir. Crit. Care Med.* **2005**, *171*, 1070–1076.
- Taleb, S.; Canello, R.; Clement, K.; Lacasa, D. Cathepsin S promotes human preadipocyte differentiation: possible involvement of fibronectin degradation. *Endocrinology* **2006**, *147*, 4950–4959.
- Tan, D. S. Diversity-oriented synthesis: exploring the intersections between chemistry and biology. *Nature Chem. Biol.* **2005**, *1*, 74–84.
- Tan, L.; Bajorath, J. Utilizing target-ligand interaction information in fingerprint searching for ligands of related targets. *Chem. Biol. Drug Des.* **2009**, *74*, 25–32.
- Tan, L.; Geppert, H.; Sisay, M. T.; Gütschow, M.; Bajorath, J. Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem* **2008a**, *3*, 1566–1571.
- Tan, L.; Lounkine, E.; Bajorath, J. Similarity searching using fingerprints of molecular fragments involved in protein-ligand interactions. *J. Chem. Inf. Model.* **2008b**, *48*, 2308–2312.
- Tavares, F. X.; Deaton, D. N.; Miller, A. B.; Miller, L. R.; Wright, L. L.; and Zhou, H.-Q. (). Potent and selective ketoamide-based inhibitors of cysteine protease, cathepsin K. *J. Med. Chem.* **2004**, *47*, 5049–5056.
- Thurmond, R. L.; Beavers, M. P.; Cai, H.; Meduna, S. P.; Gustin, D. J.; Sun, S.; Almond, H. J.; Karlsson, L.; Edwards, J. P. Nonpeptidic, noncovalent inhibitors of the cysteine protease cathepsin S. *J. Med. Chem.* **2004b**, *47*, 4799–4801.
- Thurmond, R. L.; Sun, S.; Schon, C. A.; Baker, S. M.; Cai, H.; Gu, Y.; Jiang, W.; Riley, J. P.; Williams, K. N.; Edwards, J. P.; Karlsson, L. Identification of a potent and selective noncovalent cathepsin S inhibitor. *J. Pharmacol. Exp. Ther.* **2004a**, *308*, 268–276.
- Troen, B. R. The role of cathepsin K in normal bone resorption. *Drug News Perspect.* **2004**, *17*, 19–28.



- Tully, D. C.; Liu, H.; Alper, P. B.; Chatterjee, A. K.; Epple, R.; Roberts, M. J.; Williams, J. A.; Nguyen, K. T.; Woodmansee, D. H.; Tumanut, C.; Li, J.; Spraggon, G.; Chang, J.; Tuntland, T.; Harris, J. L.; Karanewsky, D. S. Synthesis and evaluation of arylaminoethyl amides as noncovalent inhibitors of cathepsin S. Part 3: heterocyclic P3. *Bioorg. Med. Chem. Lett.* **2006c**, *16*, 1975–1980.
- Tully, D. C.; Liu, H.; Chatterjee, A. K.; Alper, P. B.; Epple, R.; Williams, J. A.; Roberts, M. J.; Woodmansee, D. H.; Masick, B. T.; Tumanut, C.; Li, J.; Spraggon, G.; Hornsby, M.; Chang, J.; Tuntland, T.; Hollenbeck, T.; Gordon, P.; Harris, J. L.; Karanewsky, D. S. Synthesis and SAR of arylaminoethyl amides as noncovalent inhibitors of cathepsin S: P3 cyclic ethers. *Bioorg. Med. Chem. Lett.* **2006b**, *16*, 5112–5117.
- Tully, D. C.; Liu, H.; Chatterjee, A. K.; Alper, P. B.; Williams, J. A.; Roberts, M. J.; Mutnick, D.; Woodmansee, D. H.; Hollenbeck, T.; Gordon, P.; Chang, J.; Tuntland, T.; Tumanut, C.; Li, J.; Harris, J. L.; Karanewsky, D. S. Arylaminoethyl carbamates as a novel series of potent and selective cathepsin S inhibitors. *Bioorg. Med. Chem. Lett.* **2006a**, *16*, 5107–5111.
- Tyndall, J. D.; Nall, T.; Fairlie, D. P. Proteases universally recognize beta strands in their active sites. *Chem. Rev.* **2005**, *105*, 973–999.
- Uhland, K. Matriptase and its putative role in cancer. *Cell. Mol. Life Sci.* **2006**, *63*, 2968–2978.
- Vasiljeva, O.; Reinheckel, T.; Peters, C.; Turk, D.; Turk, V.; Turk, B. Emerging roles of cysteine cathepsins in disease and their potential as drug targets. *Curr. Pharm. Des.* **2007**, *13*, 387–403.
- Velasco, G.; Santiago, C.; Victor, Q.; Luis, M. S.; Carlos, L. Matriptase-2, a membrane-bound mosaic serine proteinase predominantly expressed in human liver and showing degrading activity against extracellular matrix proteins. *J. Biol. Chem.* **2002**, *277*, 37637–37646.
- Vendrell, J.; Querol, E.; Avilés, F. X. Metallo-carboxypeptidases and their protein inhibitors. Structure, function and biomedical properties. *Biochim. Biophys. Acta* **2000**, *1477*, 284–298.
- Veselovsky, A. V.; Ivanov, A. S. Strategy of computer-aided drug design. *Curr. Drug Targets Infect. Disord.* **2003**, *3*, 33–40.

- Vogt, I.; Stumpfe, D.; Ahmed, H. E. A.; Bajorath, J. Methods for computer-aided chemical biology. Part 2: Evaluation of compound selectivity using 2D molecular fingerprints. *Chem. Biol. Drug. Des.* **2007**, *70*, 195–205.
- Walther, T.; Arndt, H. D.; Waldmann, H. Solid-support based total synthesis and stereochemical correction of brunsvicamide A. *Organic Lett.* **2008**, *10*, 3199–3202.
- Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.
- Wang, R.; Fang, X.; Lu, Y.; Yang, C.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.*, **2005**, *48*, 4111–4119.
- Warren, G. L.; Andrews, C. W.; Capelli, A.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- Whitesides, G. M.; Krishnamurthy, V. M. Designing ligands to bind proteins. *Q. Rev. Biophys.* **2005**, *38*, 385–395.
- Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.
- Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–1053.
- Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combin. Chem. High Throughput Screen.* **2000**, *3*, 363–372.
- Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151–1157.

- Yamashita, D. S.; Dodds, R. A. Cathepsin K and the design of inhibitors of cathepsin K. *Curr. Pharm. Des.* **2000**, *6*, 1–24.
- Yasuda, Y.; Kaleta, J.; Brömme, D. The role of cathepsins in osteoporosis and arthritis: rationale for the design of new therapeutics. *Adv. Drug Deliv. Rev.* **2005**, *57*, 973–993.
- Young, M. B.; Barrow, J. C.; Glass, K. L.; Lundell, G. F.; Newton, C. L.; Pellicore, J. M.; Rittle, K. E.; Selnick, H. G.; Stauffer, K. J.; Vacca, J. P.; Williams, P. D.; Bohn, D.; Clayton, F. C.; Cook, J. J.; Krueger, J. A.; Kuo, L. C.; Lewis, S. D.; Lucas, B. J.; McMasters, D. R.; Miller-Stein, C.; Pietrak, B. L.; Wallace, A. A.; White, R. B.; Wong, B.; Yan, Y.; Nantermet, P. G. Discovery and evaluation of potent P1 aryl heterocycle-based thrombin inhibitors. *J. Med. Chem.* **2004**, *47*, 2995–3008.
- Zakharova, E.; Horvath, M. P.; Goldenberg, D. P. Structure of a serine protease poised to resynthesize a peptide bond. *PNAS* **2009**, *106*, 11034–11039.
- Zavodszky, M. I.; Rohatgi, A.; Van Voorst, J. R.; Yan, H.; Kuhn, L. A. Scoring ligand similarity in structure-based virtual screening. *J. Mol. Recognit.* **2009**, *22*, 280–292.
- Zhao, Q.; Jia, Y.; Xiao, Y. Cathepsin K: a therapeutic target for bone diseases. *Biochem. Biophys. Res. Commun.* **2009**, *380*, 721–723.







# Eidesstattliche Erklärung

An Eides statt versichere ich, dass ich die Dissertation "*Virtual compound screening and structure activity relationship analysis: method development and practical applications in the design of new serine and cysteine protease inhibitors*" selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

Tan, L.; Geppert, H.; Sisay, M. T.; Gütschow, M.; Bajorath, J. Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem* **2008**, *3*, 1566–1571.

Crisman, T. J.\*; Sisay, M. T.\*; Bajorath, J. Ligand-target interaction-based weighting of substructures for virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 1955–1964.

Stumpfe, D.; Frizler, M.; Sisay, M. T.; Batista, J.; Vogt, I.; Gütschow, M.; Bajorath, J. Hit expansion through computational selectivity searching. *ChemMedChem* **2009**, *4*, 52–54.

Sisay, M. T.; Hautmann, S.; Mehner, C.; König, G. M.; Bajorath, J.; Gütschow, M. Inhibition of human leukocyte elastase by brunsvicamides A-C: cyanobacterial cyclic peptides. *ChemMedChem* **2009**, *4*, 1425–1429.

Sisay, M. T.\*; Peltason, L.\*; Bajorath, J. Structural interpretation of activity cliffs revealed by systematic analysis of structure-activity relationships in analog series. *J. Chem. Inf. Model.* **2009**, *49*, 2179–2189.

Stumpfe, D.\*; Sisay, M. T.\*; Frizler, M.; Vogt, I.; Gütschow, M.; Bajorath, J. Inhibitors of cathepsins K and S identified using the DynaMAD virtual screening algorithm. *ChemMedChem* **2010**, *5*, 61–64.

Frizler, M.; Stirnberg, M.; Sisay, M. T.; Gütschow, M. Development of nitrile-based peptidic inhibitors of cysteine cathepsins. *Curr. Top. Med. Chem.* **2010**, *10*, 294–322.

Sisay, M. T.; Steinmetzer, T.; Stirnberg, M.; Mauer, E.; Hammami, M.; Bajorath, J.; Gütschow, M. Identification of the first low molecular weight inhibitors of matriptase-2. *J. Med. Chem.* **2010**, *53*, 5523–35.

\* Shared first authorship

Bonn, den 08.06.2010

---

(Mihiret Tekeste Sisay)